

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

Booby trap

*Fall in seabird numbers
driven by rat infestation
harms tropical reef
ecosystem* PAGES 190 & 250

NATURE.COM/NATURE

12 July 2018 £10

Vol. 559, No. 7713

THIS WEEK

EDITORIALS

DEEP DIVE Sea-bed mining needs careful scrutiny and controls **p.152**

WORLD VIEW Science needs to acknowledge and openly discuss racism **p.153**



RESIN HOPES Mixture gives 3D printers a new, flexible look **p.155**

No place for bullies in science

High-profile allegations of bullying at a German research institute highlight the need for better systems to protect young scientists.

Picture the scene: You are an enthusiastic young scientist, with, you think, the world at your feet. You have an exciting offer to join a world-leading research institute in another country. And then, to your dismay, you find yourself in a workplace where everything feels wrong. Your supervisor intimidates you and you receive upsetting e-mails, but the institute leadership seems indifferent. You are alone in a foreign culture, and you don't know what to do. Your friends tell you to complain, but you are afraid of repercussions — and of losing the opportunity you fought so hard for. And, anyway, you don't know who to trust.

This has apparently been the situation for years for some young researchers at the Max Planck Institute for Astrophysics in Garching, Germany. Details of their struggles with alleged bullying by one of the directors — Guinevere Kauffmann — erupted in the media in the past two weeks.

According to the allegations, problems at the institute have simmered for years. The institute put in place coaching and monitoring for Kauffmann, who says: "I believe I have modified my behaviour very substantially in the last 18 months since the complaints were made." The institute also circulated an anonymous survey to young researchers, asking whether they think the problems are continuing and whether they have enough support. The results are to be presented to the institute this week, but, according to a leaked copy of the report, they show three fresh allegations of bullying against current staff, although it is not known against whom. The institute says it is investigating.

In *Nature*'s opinion, young researchers there have been let down over the years. These researchers say the institute and its parent body, the Max Planck Society — also one of the world's leading research organizations — failed to control the situation in a timely manner. It is hard to disagree.

Most scientific institutions in Germany — including the Max Planck Society — already have formal procedures in place to deal with misconduct in the lab. These focus mainly on plagiarism, fabrication and falsification. But they usually also include an ombudsman system — a supposedly independent figure who can hear complaints and weigh in. Students and staff should be able to raise allegations of bullying in this way. In this specific case, this didn't happen. Young scientists who say they were bullied found the system inadequate because no one to whom they could complain within the Max Planck system seemed to them to be truly independent. The Max Planck Institute for Astrophysics says it is already revising its code of conduct and making sure its internal mechanisms for addressing concerns are accessible. The Max Planck Society says that it will also look at the way it deals with allegations of inappropriate behaviour. It should do so urgently.

A key relationship in science is that between student and supervisor, in which the student is often entirely dependent on the supervisor and that person's support to progress in their career. Students embarking on postgraduate or postdoc years are sometimes unprepared for the demands that are placed on them. And many supervisors have only their own experiences to draw on. It's possible that what a supervisor views as

firm direction or lively banter crosses the line into bullying and abuse.

Institutions will argue — correctly — that allegations must be addressed with due process, typically with a need for confidentiality. That can and does take time. But rapid protection for those who may be suffering and, more generally, the protection of vulnerable young employees at the start of their careers must take precedence.

The issue of sexual harassment in science has received overdue and

"The protection of vulnerable young employees must take precedence."

welcome attention recently, and most observers believe the high-profile cases reported so far are the tip of an iceberg. Could bullying be equally prevalent? The data are not yet there, but it is telling that of the 300-plus responses sent in relation to a recent Editorial (*Nature* 556, 5; 2018) about poor postgraduate mental health, a significant number highlighted a

dysfunctional — or worse — relationship with their supervisor.

We will never know how many promising scientific careers around the world have been brought to a premature end because young researchers felt they could not continue to work under a bullying senior figure. But it should stop. Now. Those affected must be shown that the system will protect them if they choose to speak out. Institutions should ensure they have explicit policies in place for dealing with bullying, and, as part of that, define what constitutes bullying. And senior scientists who see colleagues behave in an inappropriate way should speak out. ■

Form is temporary

Analysis of career-long impact offers renewed hope for scientists waiting for success.

“**T**here are no second acts in American lives,” wrote F. Scott Fitzgerald, and he should know. He never quite equalled the form he found with his novel *The Great Gatsby* in 1925 — although some might say neither has anyone else.

The phenomenon of form — and how it can cluster into claimed hot streaks — is much discussed by movie buffs. Does anyone doubt that the Oscars for Martin Scorsese's 2006 film *The Departed* were belated recognition for the director's peerless streak between 1973 (*Mean Streets*) and 1980 (*Raging Bull*)? For that matter, has Robert de Niro ever topped his performances in the latter film and in Scorsese's *Taxi Driver* (1976)? It's the same story in music: the stars of Madonna, Björk and Beyoncé have all shone their brightest at particular times.

Such examples could suggest that if you haven't produced any big hits by the middle of your career, you've missed your chance. But a study published this week in *Nature* offers hope for those still waiting

(L. Liu *et al.* *Nature* <https://doi.org/10.1038/s41586-018-0315-8>; 2018). It examines the occurrence of hot streaks — runs of high-impact works — in the oeuvres of tens of thousands of film-makers, artists and scientists. It finds that most careers contain at least one relatively hot streak, and that this occurs at an apparently random stage in an individual's sequence of works.

From 'hot hands' in basketball to 'momentum' in football, folk wisdom tends to dominate discussions of form, just as it does beliefs about gamblers' winning streaks. Some will claim that 'everyone knows' artists and scientists produce their best work when they are young: Mary Shelley wrote *Frankenstein* at 19, and Jocelyn Bell Burnell was in her 20s when she discovered the first pulsar. But then, how to explain the late second blooming of novelist Philip Roth? Others place the peak of performance at mid-career, when the benefits of experience aren't yet counteracted by declining faculties — look, for example, at the musicians Ella Fitzgerald and Nina Simone.

The new analysis, which looks at crowdsourced film ratings and art auction prices, says that there is no typical career point for a hot streak. The authors argue that creative impact shows the features of 'bursty dynamics' — just like other human traits, including movement and e-mail and telephone communications (K.-I. Goh & A.-L. Barabási *EPL* **81**, 48002; 2008). This is not quite the same as saying that large or significant events happen at random; rather, their occurrence is correlated, such that the average time between successive events is smaller than random. If one occurs, another is likely to follow soon — but that sequence can't last long. That's precisely what a hot streak is.

For the 20,000 scientists included in the study, the proxy for impact was the citations of an individual's papers over the ten-year period following each paper's publication. One could quibble that some scientific papers draw most attention only decades after publication — but

that's rather rare. Hot streaks here correspond to a run of papers cited significantly more than an individual's average.

The good news is that around 90% of artists and scientists have at least one such hot streak in their career. The bad news is that it's typically not repeated: 64% of artists and 68% of scientists have only one, and more than two is very rare. F. Scott Fitzgerald was mostly right, then. And there might be little one can do to influence the matter: hot streaks do not, for example, correlate with productivity. The authors of the study make no claim that 'impact', as they measure it, is a good proxy for creativity. After all, there's still no consensus about how creativity should be defined and measured, let alone whether or how it can be cultivated and nurtured. And scientific impact goes beyond citations.

"The good news is that most careers contain a hot streak. The bad news is that it's typically not repeated."

Indeed, it would be a sad day when the intrinsic value of a work was judged by how much it can be sold for. But the disconnect between popularity and worth perhaps goes to the heart of what to make of these findings. To use economics terminology, are the dynamics of success endogenous — driven by the fluctuating inspiration of the creator, say — or exogenous, produced by the vicissitudes of the marketplace? It's tempting to imagine a bit of both: that the creator suddenly finds he or she has tapped into the zeitgeist — only to discover, a little further down the line, that the world has moved on.

Maybe the most appealing message, however, is that the dynamics of science are no different from those of the arts: success in both depends on a resonance between the individual's imagination and the shifting moods and desires of the audience. ■

Depth charge

Nations must weigh the environmental costs before mining minerals on the ocean floor.

The biggest deep-sea mining operation so far was a cold-war ruse. In 1974, the US Central Intelligence Agency launched an elaborate operation to recover a Soviet submarine northwest of Hawaii, under the cover of a commercial venture to mine manganese nodules located on the sea floor. The spooks got a piece of the submarine but left any valuable minerals in the area for future prospectors.

Despite mounting interest in such sea-bed resources, little has happened since. But the world's first commercial mining venture in the deep sea seems only a matter of time. Estimates vary, but optimists claim there could be huge untapped piles of precious metals, including gold and silver, down there. Several countries, including Papua New Guinea, Japan and South Korea, are pursuing sea-bed mining in their territorial waters, and interest in international waters is on the rise, as well. The International Seabed Authority (ISA), which was established in 1994 under the United Nations' Law of the Sea and regulates activities that occur outside national jurisdictions, has already issued 28 exploration permits to entities in 20 countries.

At a meeting next week, the ISA's council will once again discuss a possible process to permit, monitor and end mining operations, which will eventually cover more than half of the ocean floor — collectively known as 'the Area'. The ISA began deliberations in earnest in January 2017, and observers think that the regulations could be formalized as early as 2020. Many of the details must still be ironed out.

The impact of the regulations will necessarily be limited: 29 countries, including the United States, have yet to either sign or ratify the Law of the Sea, and so are not technically bound by the ISA's authority. And within national waters, each country must craft sensible regulations. But the

ISA could help to create best-practice policy that will aid governments in regulating domestic activities.

For a time, it looked as if the first commercial project could be in Papua New Guinea, where the Canadian firm Nautilus received its first lease in 2011. The site is located at a depth of 1,600 metres in the Bismarck Sea, between the islands of New Britain and New Ireland, but the company's quest for copper and gold deposits has encountered numerous legal and economic challenges. It remains unclear when or whether operations will move forward.

The slow pace of development has given a little breathing room to scientists who are scrambling to understand the impacts of sea-bed mining. Already, researchers have raised a host of potential problems. The most obvious is that scouring the ocean floor will destroy a landscape — and an ecosystem — that is unlikely to recover quickly: after all, it took millions of years for these mineral deposits to form.

But the impacts could stretch well beyond the immediate mining area. The dredging and collection process create underwater plumes of particles, and any material left over after the initial processing on ships could be released back into the ocean. That material will spread and eventually sink, potentially affecting life at all ocean depths.

The ISA's priority now must be to create strong environmental safeguards. In particular, regional environmental management plans must be put in place before any mining occurs. This is where research comes in: experts must determine how much of the ocean floor should be set aside, and where, to preserve biodiversity. And it can help to explain how dense particle plumes can become, and how far they should be allowed to spread. Researchers worldwide are investigating these issues.

Governments must look at mining and minerals holistically. Where possible, the goal must be to reduce consumption and increase recycling, through more-systematic recovery efforts and better product designs. But demand for metals is unlikely to abate soon. And mining on land comes with its own environmental impacts. It could even be that venturing into the sea — if done properly, and with the correct safeguards — makes more sense than digging yet another ugly hole in a sensitive terrestrial environment. But we can't say so yet. ■



Lab heads should learn to talk about racism

Senior academics must step up and take the lead in discussing intolerance, says Devang Mehta.

Last month, anti-Asian graffiti was painted in residences on the campus of my PhD alma mater, the Swiss Federal Institute of Technology (ETH) Zurich, and Asian students' work was vandalized with racist slogans. That same week brought allegations that a leading astrophysicist at the Max Planck Institute for Astrophysics in Garching, Germany, had used racist language towards trainees, among other bullying. (The astrophysicist has defended her behaviour, and says her comments were distorted and taken out of context; see page 159.)

When blatantly racist incidents occur in our universities, we academics usually prefer not to address them. We leave their handling to university administrators, who tend to deal only with the most serious cases, frequently long after they have happened. In my experience, scientists often do a poor job of recognizing and dealing with racism in our workplaces. In fact, several colleagues I spoke to while writing this article expressed scepticism that racial bias even exists in the often highly international scientific work environment. This blindness to the issue keeps us from addressing racism within the close-knit structures of academic labs.

My own experiences pale in comparison to others', but are still worth recounting. I came to Europe as a graduate student from India in 2012, just as terrorism and the refugee crisis were sparking a sharp increase in anti-immigrant rhetoric. However, working in incredibly diverse labs, I felt largely insulated.

This changed when a colleague asked me to tell a Muslim colleague off for having an untidy workbench because 'they' respond better to male authority. All I could do was stare, dumbstruck. In another instance, when asked about supporting diversity in a meeting with students, a European professor laughingly admitted to not hiring Asian researchers because he found 'them' difficult to work with. And I've heard many scientists casually dismiss all published papers from labs in certain countries as bad science, in the presence of students from those very countries.

I deeply regret that during my PhD I did not talk about these experiences with my supervisors. By not doing so, I denied them the opportunity to learn from and address my concerns in the manner in which I'm now confident they would have done.

Why didn't I work up the courage to report my concerns? I didn't want to rock the boat. Like many scientists from ethnic-minority groups, I was an immigrant lacking the social and economic safety nets that citizens enjoy. It was so much easier to put my head down and race towards that PhD.

Although official policies such as institutional codes of conduct and instruments of redress for serious offences are essential, individual principal investigators (PIs) also need to model the sort of communication that is lacking today. If the reluctance of junior researchers like me to talk about racism is regrettable, the silence, and hence complicity,

of senior faculty members is unconscionable. Scientists, as a community, must practise having tolerant conversations about intolerance, unconscious bias, unfair power structures and a friendlier workplace for everyone. And that just isn't happening: both the targets of and witnesses to microaggressions worry that they are reading too much into certain actions. Relevant incidents rarely reach the attention of PIs.

The lead must come from the top — from PIs, deans, provosts. The first step could be something as simple as showing a willingness to hear about racism and intolerance from students and employees. I have asked around, and I have not heard of a single instance in which a lab head, of any race or ethnicity, male or female, held a lab meeting or sent a welcome e-mail explicitly recognizing that these are real problems they are willing to discuss. I write publicly about these topics, but I find it hard to even imagine raising racism or inequality with supervisors in face-to-face meetings unless they first signalled an openness to talk about them.

It's not easy to call out colleagues over racist comments or intolerant behaviour, but we must. For inspiration, I sometimes consider the universal ethical code for scientists devised in 2007 by David King, then the UK government's chief scientific adviser, which requires high standards of integrity for evidence and society (go.nature.com/2u7ytdt). And guidelines exist for essential conversations, for example those from the Massive Science Consortium, a group of more than 300 young scientists of which I'm a member (go.nature.com/2tsauch). One tenet is "assume good intentions and forgive". Talking about race can lead to people feeling persecuted, fairly or unfairly, and forgiveness is needed to move on from a confrontational or racist incident. (Assuming, of course, that the incident was minor, and apologies were offered.)

Another guideline is "step back and step up". This asks privileged individuals to make sure they don't dominate a discussion, and to listen to contributions from minorities and less powerful groups.

Perhaps the most important guideline is "speak and listen from personal experience". In other words, do not instinctively question the validity of someone else's experience; this happens so often with women and minorities. It is especially apparent when institutions reflexively defend the accused. It is up to tenured professors to protest and demand more introspection from their employers and employees.

Fundamentally, tackling racism and intolerance in science requires an acknowledgement from us all that it exists. I call on senior scientists to speak up and to invite others to do so. ■

SCIENTISTS, AS A COMMUNITY, MUST PRACTISE HAVING TOLERANT CONVERSATIONS ABOUT INTOLERANCE.

Devang Mehta is a postdoctoral fellow in the Laboratory of Plant Genomics at the Department of Biological Sciences, University of Alberta in Edmonton, Canada.
e-mail: devangmehta@ualberta.ca; @drdevangm

POLICY

Gender policy

Germany's main grant-giving agency, the DFG, has unveiled new policies to improve standards around gender equality in funding proposals. At its annual meeting last week, the agency adopted a requirement that, when relevant, grant applicants outline any sex- and gender-related aspects of their research in their proposals. Biomedical researchers, for example, will need to state whether they intend to use male or female human cells in research projects. The US National Institutes of Health implemented similar policies in 2014. The move is part of the DFG's qualitative gender-equality strategy, adopted in 2017. The agency says that it will further examine its funding programmes and review criteria to identify obstacles to gender inclusion and diversity in research.

Funding autonomy

Chinese Premier Li Keqiang says the country will ease restrictive rules governing how research funds are used and cut red tape for researchers. In the past, Chinese researchers have complained that they cannot use grants to pay accountants to manage their finances or to purchase urgently needed equipment, without going through a lengthy bidding process. They say that this puts them at a disadvantage relative to researchers elsewhere. Addressing a meeting of China's high-powered state council on 4 July, Li vowed to remove such restrictions and give researchers new autonomy to increase their productivity and the vitality of their research. "We must remove the restraints on researchers



ASAHI SHIMBUN VIA GETTY

Deadly floods inundate western Japan

Torrential rains across western Japan have caused flooding and landslides that have killed more than 100 people. More than 5 million others were ordered or advised to evacuate. The rains started on 5 July and are associated in part with a typhoon. They were the heaviest seen in Japan for decades. Houses collapsed or were buried by landslides, and roads and train routes were closed. Prime Minister Shinzo Abe cancelled a trip to Europe and the Middle East to oversee the disaster response.

as soon as possible so that they can devote themselves to their research," he said.

PEOPLE

Environment chief

Scott Pruitt, the embattled administrator of the US Environmental Protection Agency (EPA), resigned on 5 July amid a series of ethical and spending scandals. He also provoked controversy over his aggressive agenda to roll back health and environmental regulations. Pruitt sought to reverse a series of rules aimed at reducing greenhouse-gas emissions from power plants, oil and gas operations and other industries. He also changed how the agency approaches scientific research, banning scientists with EPA funding from serving on agency advisory boards

and proposing limits on the kinds of research used to justify regulations. Andrew Wheeler, the EPA's deputy administrator, will be its acting chief until the US Senate can confirm a new nominee.

Chief scientist fired

Doug Ford, the premier of Ontario, Canada, fired the province's first chief scientist on 3 July. The previous administration appointed biomedical engineer Molly Shoichet to the post in November 2017, with the goal of advancing science and innovation throughout Ontario. Shoichet also advised the previous premier on scientific matters. Since Ford's election on 29 June, he has announced his intention to cancel a cap-and-trade carbon-pricing programme on which the

province collaborates with neighbouring Quebec and with California. He has also instructed ministries to cancel subscriptions to publications including newspapers, magazines and trade journals. Some worry that this could include scientific journals.

EVENTS

Science council

Two organizations have merged to create the International Science Council (ISC), a non-governmental body that will provide scientific advice to international organizations. The ISC, which launched on 5 July in Paris, unites the International Council for Science and the International Social Science Council, founded in 1931 and 1952, respectively. The

PETER JENNISSENS
ISC's members include 40 international organizations and more than 140 national and regional bodies, including academies and research councils. Its activities include convening experts to advise on issues of scientific and public importance to organizations such as the United Nations on topics such as biodiversity, climate change and the UN's sustainable development goals. The ISC elected Daya Reddy, a mathematician at the University of Cape Town, South Africa, as its first president. In his acceptance speech, Reddy called for the ISC to be inclusive of all regions and to involve early-career scientists in its work.

Space fragment

Meteorite hunters have retrieved a fragment of the space rock that lit up the sky over Botswana last month. Astronomers spotted the incoming asteroid on 2 June, just hours before it hit Earth and triggered a massive ground search around the impact area. Lesedi Seitshiro, a geologist at the Botswana International University of Science and Technology in Palapye, spied the 18-gram meteorite (pictured) in Botswana's Central Kalahari Game Reserve. The University of Helsinki, which participated in the search, announced the finding on 6 July. This feat is a rare example of scientists



tracking an asteroid from before impact through to when it winds up as fragments on the ground.

RESEARCH

Papers retracted

The prestigious medical journal *The Lancet* retracted two articles co-authored by disgraced thoracic surgeon Paolo Macchiarini, after an investigation by his former institute found him responsible for scientific misconduct. The publications, one a research paper published in 2011 and the other a review published in 2012, were retracted on 7 July and relate to an experimental transplant technique that involved implanting artificial windpipes seeded with stem cells into a patient. The surgery, carried out in three people between 2011 and 2013, failed. An investigation released on 25 June by the Karolinska

Institute in Stockholm, where Macchiarini was a visiting faculty member, found that he and six co-authors were responsible for misconduct in six studies, two of which are the *Lancet* papers. The institute found that the articles contained "fabricated and distorted descriptions of the patients' conditions before and after the operations", as well as other inaccuracies. The president of the institute, Ole Petter Ottersen, requested that all six studies be retracted.

PUBLISHING

Sweden cut off

Researchers in Sweden have lost access to the latest articles published in Elsevier journals, after negotiators failed to agree on a new contract. Talks between the academic publisher and the Bibsam Consortium, which brokers journal-subscription deals on behalf of 85 Swedish

institutions, have founded over the consortium's demands that the next agreement cover fees charged to researchers to publish their work under open-access terms. The previous deal ended on 30 June. Scientists cannot access papers published after this date, but material published between January 1995 and 30 June remains available. Elsevier said that it is open to talks that would "find a sustainable solution in support of Swedish research and its open-access ambitions". Similar negotiations have broken down in Germany, where the DEAL consortium, which represents German institutes, said on 5 July that it was suspending talks with Elsevier. DEAL has since 2016 been negotiating with the publisher over the cost of making newly published research open access. Around 185 German institutions now have no contract with Elsevier, but the publisher has mostly granted them continued access to its journals while talks continue.

Open-science plan

The French government has unveiled a plan for making publicly funded research open access. Research minister Frédérique Vidal said on 4 July that it should be mandatory for scientists who carry out research supported by government money to publish their results in open-access journals or repositories. The government will invest €5.4 million (US\$6.3 million) in 2019 towards developing the plan, and another €3.4 million per year until the end of the government's term in 2022. The plan will help to develop an existing national open-research repository called HAL, and will allow France to contribute to open-science projects across Europe. The proposal also includes cash to transfer some paywalled research content to open-access repositories.

NATURE.COM

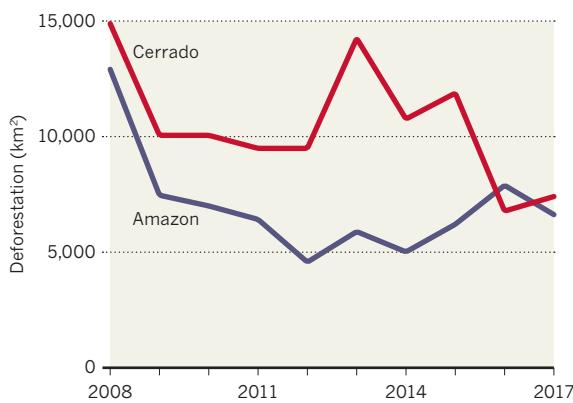
For daily news updates see:
www.nature.com/news

TREND WATCH

Deforestation in the savannah bordering the Brazilian Amazon spiked by 9% last year. An estimated 7,408 square kilometres of land were cleared for agriculture in the Cerrado region — a sharp drop from several years ago but still more than the total deforestation in the Amazon in the same period. The Cerrado is less than half the size of the Amazon and has already lost roughly half of its original landscape, the Amazon Environmental Research Institute in Brasilia reported on 28 June.

CERRADO STILL THREATENED

Deforestation in the savannah region bordering the Brazilian Amazon has risen slightly, but roughly half of the biome has already disappeared.



NEWS IN FOCUS

POLITICS Researcher figures stress need for immigration reform ahead of Brexit **p.160**

MARIJUANA Coming to a lab near you: genetically engineered cannabis **p.162**

CHEMICALS Machine learning enlisted to predict toxicity **p.164**



NEUROSCIENCE A bat researcher investigates how brains navigate in 3D **p.165**

© GHOST WRITER VIA WIKIMEDIA COMMONS/CC BY 4.0



The Max Planck Institute for Astrophysics in Garching, Germany, is carrying out a survey to find out whether people are being bullied.

GERMANY

Max Planck bullying controversy intensifies

Astrophysicist defends her behaviour and her institute steps up its response.

BY ALISON ABBOTT

Asaga concerning allegations of abuse at the Max Planck Institute for Astrophysics (MPA) in Garching, Germany, is coming to a head. An MPA director accused of bullying has for the first time spoken out to defend herself. And the prestigious Max Planck Society, which funds the MPA, is investigating fresh allegations of bullying and sexual harassment at the institute following an anonymous survey of its young scientists — it is not clear whom these new allegations concern.

The difficulties at the MPA first surfaced publicly in an article in the news magazine *Der Spiegel* in February. The article detailed accusations of bullying of graduate students and postdocs by a director at an unnamed Max Planck institute in Bavaria. A news report by BuzzFeed Germany on 27 June then named the director as astrophysicist Guinevere Kauffmann, and added further details — including allegations of racist comments.

Now, Kauffmann has corresponded with *Nature* to explain the alleged behaviour. “I am not a racist. I am half Chinese and half

German-Jewish,” she writes. “Because of my mixed-race background, I am very interested in cultural differences between people and I regret very much that my comments have been taken out of context and distorted,” she adds.

“Regarding ‘bullying’ — I am of the generation that was subjected to very high-pressure supervision. I realize that this has now become unacceptable. I believe I have modified my behaviour very substantially in the last 18 months, since the complaints were made.”

One MPA graduate student who spoke to *Nature* on the condition of anonymity said ▶

► that bullying had been a major disruptive force at the institute, causing, in the student's opinion, at least two young researchers to leave their positions prematurely.

Kauffmann says: "Our procedures for evaluating graduate students and providing honest feedback are still not uniform enough, in my opinion. Approaches vary greatly, from 'keep quiet if the student is not doing well and let him/her sink in the exam or job market', to 'attempt to steer towards a successful career with all your strength'. I believe I fall into the latter category. Nobody I had trouble with ended up quitting astronomy."

MEASURES FOR IMPROVEMENT

The MPA leadership learnt of the bullying allegations in 2016, when the institute's external scientific advisory board described a complaint from young scientists, says Eiichiro Komatsu, a director at the MPA. Komatsu, who was managing director at the time, says that he and his colleagues responded immediately, and provided coaching for Kauffmann, who also agreed to daily monitoring.

Software engineer Andressa Jendreieck, who was a graduate student at the MPA between 2011 and 2014, told *Nature* that for

many years, young researchers were afraid to make complaints and believed that there was no independent person they could turn to.

In its report to the MPA leadership, the external board noted that "there is no effective mechanism for individuals at the MPA to file formal complaints to the Max Planck Society if they have been treated inappropriately by other members of the institute", says Komatsu.

In response to the February article in *Der Spiegel*, the MPA conducted an anonymous survey of its young scientists. It asked about their experiences of bullying or sexual harassment at the institute, among other things.

The results, which were due to be presented at the institute on 13 July but have now been leaked, show that the MPA sent the survey to 120 master's students, PhD students and post-docs, and that just over half responded. Three report that they were bullied and two report that they were sexually harassed. It is not clear whether these new accusations are related to the earlier allegations, nor whom they concern.

The Max Planck Society says that it has commissioned an independent law firm to investigate the new allegations. "We need to clearly define these allegations in order

to assess the severity of the incidents and to intervene accordingly," says the society's press officer, Christina Beck.

Beck says that contact details of the law firm will be sent to MPA staff in the coming weeks — and that scientists will be able to speak to the firm in full confidentiality. The firm will report its conclusions to the MPA leadership. Beck hopes that those affected will take advantage of the independent mechanism to report their allegations.

The unnamed graduate student who spoke to *Nature* says that researchers would probably engage with such a process. But the student also notes that, in their opinion, confidence in the Max Planck leadership has slipped because its responses in 2016 came too late, and were not tough enough.

The allegations at the MPA come in the wake of separate complaints by scientists at the Max Planck Institute for Biological Cybernetics about how the society is handling animal-welfare charges against a leading neuroscientist. Beck says that the institutes are independent of the society's general management, which only advises the institutes' leaderships and checks administrative procedures. ■

RESEARCHER MOBILITY

Scientists call for migration reform before Brexit

Figures on foreign-researcher mobility highlight need for UK policy change.

BY ELIZABETH GIBNEY

The UK immigration system may need to process tens of thousands more visas for scientists each year if European Union citizens lose their special immigration rights after Brexit, figures obtained by *Nature* suggest. The numbers underscore the urgent need for reform of the rules governing immigration by researchers — a topic that a parliamentary group has been investigating since May.

Immigration data gathered by *Nature* also highlight that the current system is not working well for scientists who come from outside the EU, irrespective of Brexit. One type of visa — called Tier 1 Exceptional Talent, and designed to attract leaders and emerging leaders from overseas, largely in the sciences and engineering — is vastly underused, with fewer than half of a possible 1,000 visas taken up last year. And non-EU researchers already often struggle to get visas for short visits for conferences and collaboration.

Despite recent tweaks to immigration rules in favour of researchers, many scientists see Brexit as an opportunity for further, much-needed reform to the entire system for highly skilled workers. "Maybe when the dust settles we can get a system that's better for those coming from all over the world," says Richard Catlow, foreign secretary at the Royal Society in London.

SYSTEM CHANGE

Immigration data requested by *Nature* from the UK Home Office under the Freedom of Information Act show that the United Kingdom approved visas for about 20,000 academic researchers and non-academic PhD-level research professionals from outside the EU in the 2016–17 academic year (see 'Researcher mobility after Brexit').

Because EU nationals currently have automatic rights to work in and travel freely to the

"The system would have to deal with an approximate doubling in capacity."

United Kingdom, comparable figures for the number of European researchers entering the country do not exist. However, data from the Higher Education Statistics Agency show that over the same period, UK universities hired about 10,500 researchers who were EU nationals: 5,760 full-time academics and 4,835 post-graduate research students. And calculations by *Nature*, based on migration data and annual labour surveys, suggest that each year, thousands more EU citizens take up research roles in UK industry, charities and government.

Neither count includes visitors coming to Britain on short trips, such as to attend conferences or for collaboration meetings. But data from the UK Office of National Statistics shows that in 2016, EU residents made more than three times as many business visits — which would include short, scientific trips — as did citizens from the rest of the world combined.

After Brexit, many of these thousands of EU researchers coming to the United Kingdom are likely to need some form of visa (the country

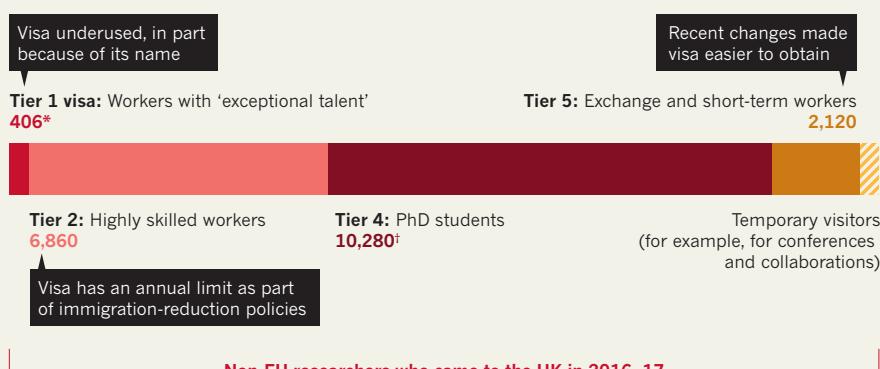
RESEARCHER MOBILITY AFTER BREXIT

The UK immigration system may need to process tens of thousands more visas each year for European Union scientists if EU citizens lose their rights to live and work in the United Kingdom after Brexit, suggest figures obtained by *Nature*.

 Precise data are not readily available

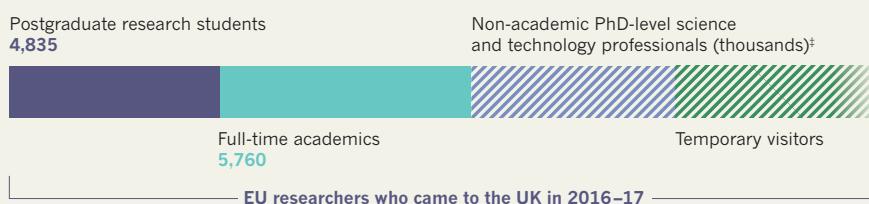
RESEARCHERS FROM OUTSIDE THE EU

Non-EU scientists enter the United Kingdom on a variety of visas, with 19,666 visas for academic and non-academic PhD-level researchers approved in the last academic year. The system currently limits how many 'highly skilled' workers can come in.



RESEARCHERS FROM THE EU

Some 10,500 EU researchers started long-term positions in UK academia last year, according to available data, but the true number of incoming EU researchers is likely to be thousands higher when those coming to work in industry, charities and government are included.



*Applicants for Tier 1 visas are endorsed by eminent organizations. Data shown are for endorsements rather than approvals, but most applicants who are endorsed receive visas, according to the organizations. †Higher Education Statistics Agency data on new non-EU starters in postgraduate research degrees used as a proxy for Tier 4 visas. ‡Order of magnitude estimate based on the 3.5% of EU arrivals in 2014-17 who were listed as working in PhD-level jobs in the 2017 Labour Force Survey. Tier 3 does not exist.

leaves the EU in March 2019, but EU nationals retain their current rights until the end of 2020). "These numbers tell us that if you were going to expand the non-EU system, it would have to deal with an approximate doubling in capacity," says Sarah Main, executive director of the London-based Campaign for Science and Engineering (CaSE).

SPECIAL MEASURES

Main and Catlow are two of the many representatives of science organizations and researchers who have submitted evidence to the parliamentary inquiry, set up by the House of Commons' science and technology committee to address

the issue of researcher mobility in the wake of Brexit. The government has pledged to restrict migration from the bloc but it is yet to publish its long-awaited overall plans for immigration after the departure. The committee hopes its recommendations will inspire immigration solutions that would best serve science, says Norman Lamb, the member of parliament leading the inquiry. Many people would be alarmed by the idea of fitting EU citizens into the existing system without any adjustment, he says.

Research organizations responding to the inquiry say that one straightforward option for reform is to permanently remove a cap on Tier 2 visas, which are given to highly skilled

workers and are the channel through which most non-EU scientists come to work in the United Kingdom. The cap, part of government efforts to reduce overall immigration, is unpopular irrespective of Brexit: between December and March, for instance, it meant that 3,500 eligible technicians, engineers, and science and technology professionals were refused visas.

Scrapping the cap has growing support in the government, says Jonathan Portes, an economist at Kings College London who is leading a project on how Brexit is likely to affect immigration.

And, on 6 July, the government announced immediate changes to a different visa route, called Tier 5, that will expand the range of institutions that can sponsor researchers and technicians hired on placements of up to two years. "It's another positive visa change," says Main.

Organizations say that overhauling the entire system for highly skilled workers is a priority—but some worry that it is unlikely to happen by the end of 2020. As a result, many have proposed mechanisms to the inquiry that would cater for EU nationals in the years directly after 2020, to address the immediate problem of researcher mobility. These mechanisms, if implemented successfully, would give the government more time to work on broader reform and could form the basis of a better overall system.

One of these temporary measures, suggested by the biomedical-research charity the Wellcome Trust, would be to extend visa-free travel for conferences and collaboration visits—already available to US and Canadian nationals—to EU citizens for a period after 2020. For longer stays, the trust has suggested trialling a 'science visa' for EU nationals; a body such as UK Research and Innovation, the national science funder, would endorse applicants with job offers to speed up their applications. Countries including Canada, France and Singapore have researcher-specific visas, the committee heard.

Negotiations about several aspects of the UK-EU relationship after Brexit are ongoing and it is still not clear through which avenue any provisions for researcher mobility might be made. The government is expected to release its immigration plans in September. But the government has also made it clear that it wants a special pact on science and innovation with the EU after Brexit—to allow the United Kingdom to access EU research money—and this could specify special measures for researchers. ■

MORE ONLINE

TOP NEWS



Scientists applaud resignation of US environment chief go.nature.com/2uka7jm

MORE NEWS

- Contagious cancer could have wiped out America's first dogs go.nature.com/2zr8zfx
- Hybrid rhino embryos created to stop extinction go.nature.com/2kmjazm
- Hunt for dark matter turns to ancient minerals go.nature.com/2m4fbmg

NATURE PODCAST



Rats and coral reefs; charting successful career streaks; and Cape Town's water crisis nature.com/nature/podcast



Medical marijuana grown the old-fashioned way in a Canadian facility.

to work on cannabis are restricted to one main supplier. The only facility in the United States certified to provide them with cannabis and its extracts is the University of Mississippi in Oxford. Scientists can also request permission to study a small number of synthetic cannabinoids from pharmaceutical companies, but some say that these sources are too limited or expensive to be of use.

“It takes a great deal of endurance to study cannabinoids,” says Ziva Cooper, a neuroscientist at Columbia University in New York City. In February, Cooper and her colleagues reported that people who smoked marijuana sourced from the University of Mississippi, and who took half the typical dose of oxycodone, experienced similar pain relief to people who took only the full opioid dose (Z. D. Cooper *et al. Neuropsychopharmacology* <http://doi.org/crw2>; 2018). To find out whether this combination might enable doctors to prescribe lower opioid doses — and therefore reduce the risk of opioid addiction — Cooper would like to conduct a larger trial. But she has yet to get approval for the study because of restrictions on marijuana research.

RICHARD LAUTENS/GETTY

FILLING DEMAND

If legal barriers fall, scientists will want to explore high-quality cannabinoids produced through various means. Marijuana compounds made using genetically engineered bacteria and yeast might help to meet the demand.

Kevin Chen, head of the biotechnology company Hyasynth Bio in Montreal, Canada, says that researchers have expressed interest in buying the company’s engineered CBD as soon as it scales up production. In May, a Canadian medical-cannabis company, Organigram in Moncton, announced its intention to invest Can\$10 million (US\$7.6 million) into Hyasynth to help boost manufacturing.

Another Canadian company, InMed Pharmaceuticals in Vancouver, is refining the production of rare cannabinoids in the bacterium *Escherichia coli*. Extracting useful amounts of these potentially beneficial compounds from plants is unrealistic because they occur at very low levels, says Samuel Banister, a chemist at the University of Sydney in Australia. “For minor cannabinoids,” he says, “there is a huge need for synthetic biology.”

If the DEA decides to remove only Epidiolex from the list of schedule 1 substances, and not CBD generally, researchers in the United States might not be able to take advantage of these companies’ products. Instead, the substances will flow to laboratories in Canada, where medical and recreational marijuana will be legal as of 17 October. Or, research might sprint forward in Germany and the Netherlands, where Kayser says scientists face few barriers to studying cannabis. Anticipating a demand, he has a patent pending in Europe on the production of cannabinoids in engineered yeast. ■

MEDICAL RESEARCH

Transgenic pot could soon hit labs

Scientists might be able to draw from new sources of cannabis compounds for research.

BY AMY MAXMEN

Legal hurdles to exploring marijuana’s medicinal properties might soon fall in the wake of the US Food and Drug Administration’s (FDA) first approval of a cannabis-derived drug.

On 25 June, the FDA announced its approval of Epidiolex — a treatment for epileptic seizures that is based on a cannabis compound called cannabidiol (CBD). The US Drug Enforcement Administration (DEA) has until 24 September to re-classify Epidiolex so that it’s legal for doctors across the country to prescribe it. Many researchers hope that the agency will re-classify CBD itself, instead of just Epidiolex, so that they can more easily study this non-psychadelic component of marijuana.

Now that the FDA has approved Epidiolex, “we have a clear recognition that this plant has more potential than people credited it for, and that has reverberations that are scientific as well as legal”, says Daniele Piomelli, director of a new centre for cannabis research at the University of California, Irvine. At the very least, he says, the DEA ought to grant researchers an exemption permitting them to study CBD

— especially now that people consume it and other cannabis compounds, known as cannabinoids, in states where marijuana is legal. At this point, the limits on research seem irrational, he adds.

Lessening restrictions on the study of CBD would also be good news for biotech startups that have been producing cannabinoids through genetic engineering. These products could be purer and more affordable than those obtained through older methods of extraction from marijuana plants or chemical synthesis.

“It’s a biochemical gold rush right now,” says Oliver Kayser, a bioengineer at the Technical University of Dortmund in Germany.

Thirty states and the District of Columbia have now legalized medical marijuana. But the plant and its compounds are still illegal under US federal law, consigned to the most restricted category of substances — schedule 1. Only the few researchers who sink the time and money into complying with federal rules for handling illicit substances can work on cannabis. Far fewer barriers block research on drugs in less-restricted categories, such as oxycodone (OxyContin) — a commonly prescribed opioid — or cocaine and ketamine.

But even researchers who have permission

ARTIFICIAL INTELLIGENCE

Software improves toxicity tests

Machine learning trumps animal testing for many chemicals.

BY RICHARD VAN NOORDEN

Machine-learning software trained on masses of chemical-safety data is so good at predicting some kinds of toxicity that it now rivals — and sometimes outperforms — expensive animal studies, researchers report.

Computer models could replace some standard safety studies conducted on millions of animals each year, such as dropping compounds into rabbits' eyes to check if they are irritants, or feeding chemicals to rats to work out lethal doses, says Thomas Hartung, a toxicologist at Johns Hopkins University in Baltimore, Maryland. "The power of big data means we can produce a tool more predictive than many animal tests."

In a paper published in *Toxicological Sciences* on 11 July, Hartung's team reports that its algorithm can accurately predict toxicity for tens of thousands of chemicals — a range much broader than other published models achieve — across nine kinds of test, from inhalation damage to harm to aquatic ecosystems (T. Luechtefeld *et al.* *Toxicol. Sci.* <http://doi.org/crw4>; 2018).

The paper "draws attention to the new possibilities of big data", says Bennard van Ravenzwaay, a toxicologist at the chemicals firm BASF in Ludwigshafen, Germany. "I am 100%

convinced this will be a pillar of toxicology in the future." Still, it could be many years before government regulators accept computer results in place of animal studies, he adds. And animal tests are harder to replace when it comes to assessing more-complex harms, such as whether a chemical will cause cancer or interfere with fertility.

COMPUTER SAYS: NOT TOXIC

Industry and academia have used computer models for decades to predict toxicity. These models typically incorporate a molecule's chemical structure, an understanding of how it might react in the body and data from animal tests or *in vitro* studies. Companies also infer the toxic effects of untested substances by comparing them with other structurally or biologically similar compounds whose effects are known — a method known as read-across. But regulators set a high bar for accepting these methods and tend to ask for animal studies instead, Hartung and other toxicologists say.

To improve the software, Hartung's team created a giant database with information on roughly 10,000 chemicals based on some 800,000 animal tests. These data

"I am 100% convinced this will be a pillar of toxicology in the future."

were originally collected by the European Chemicals Agency (ECHA) in Helsinki as part of a 2007 law known as REACH (registration, evaluation, authorization and restriction of chemicals), which requires companies to register safety information for most chemicals marketed in the European Union. As of May 2018 — the closing date for registrations — the agency had received information on more than 20,000 substances.

The ECHA makes those data public, but not in a format that allows computers to analyse them easily. So, in 2014, Hartung's team extracted the available data into a machine-readable database. Using the read-across method, Hartung's software compares a new chemical to known, closely related compounds and assesses the probability of toxic effects by reference to their established properties. Effectively, says Hartung, the software mimics how a toxicologist would size up a new chemical, but in automated fashion.

Hartung's database analysis also reveals the inconsistency of animal tests: repeated testing of the same chemical can give different results, because not all animals react in the same way. For some types of toxicity, the software therefore provides more-reliable predictions than any individual animal test, says Hartung, who has also commercialized his work.

Other researchers and firms are developing machine-learning algorithms, too, although they have not published papers about their work. And chemical-safety agencies are paying close attention. In April, the Interagency Coordinating Committee on the Validation of Alternative Methods, which is developing methods to replace animal-safety testing on behalf of 16 US government agencies, invited dozens of academic and commercial research groups to the National Institutes of Health (NIH) in Bethesda, Maryland. There, each team used its own software to predict 'lethal-dose' toxicity for 40,000 chemicals previously tested on rats.

Combining the best software (including Hartung's) produced a consensus computational model that "performed just as well as the animal tests", says Nicole Kleinstreuer, who coordinated the exercise and develops alternative toxicity-testing methods for the US National Toxicology Program in Durham, North Carolina. Later this year, the US Environmental Protection Agency (EPA) plans to release the consensus model online for free download.

In the EU, the ECHA has encouraged companies to avoid animal tests by using read-across and methods based on analysis of lab cells where possible, says Mike Rasenberger, head of computational assessment at the agency.

The new paper is part of "a good initiative", Rasenberger says, but "scientifically, there is a lot of work to be done". He adds: "No one wants animal tests, but we can't yet do all toxicology with a computer." ■



Computer programs can, in some cases, predict chemical toxicity as well as tests done on rats.

GENE EDITING

Gene drives tested in mammals for first time

Technology worked inconsistently in mice.

BY EWEN CALLAWAY

A controversial technology that can alter the genomes of entire species has been applied to mammals for the first time. In a preprint published on 4 July, researchers describe developing 'gene drives' in mice using CRISPR gene editing — and say that the technique works inconsistently in the animals.

Gene drives ensure that more of an organism's offspring inherit a certain, 'selfish' gene than would happen by chance, allowing a mutation or foreign gene to spread quickly through a population. They occur naturally in some animals, including mice. But the CRISPR–Cas9 gene-editing tool has allowed the creation of synthetic gene drives designed to eliminate problem species by, for instance, making offspring infertile. They have already been created in mosquitoes in the lab, as a potential malaria-control strategy,

and researchers have suggested that the technology could help to kill off rodent pests. The technique has attracted controversy — and even a failed attempt to ban its global use — because, if released in the wild, organisms carrying gene drives might be hard to contain.

The researchers behind the latest study, led by Kim Cooper, a developmental geneticist at the University of California, San Diego, say their goal was to create a test bed for the technology in mammals (H. A. Grunwald *et al.* Preprint on bioRxiv <http://doi.org/crw3>; 2018). Working in mouse embryos, they biased the inheritance of a mutation that gives mice all-white coats. The mutation was not always copied correctly, and the process worked only in female embryos. The team estimated that this could lead to a mutation being transmitted to 73% of a female mouse's offspring, on average, instead of the usual 50% for most genes. Cooper declined to

comment on the work, because it has not yet been published in a peer-reviewed journal.

There is an indication that the technology could work, but the study is also sobering, says Paul Thomas, a developmental geneticist at the University of Adelaide in Australia. "There is a lot more to do before you could consider gene drives for a useful tool for population control of rodents," he says. ■

CORRECTIONS

Due to a misunderstanding from our reporter, the Editorial 'Military work threatens science' (*Nature* **556**, 273; 2018) incorrectly implied that the Astronomical Society of Japan formally advocates the use of funding from the country's military research fund to support academic researchers. While some members agree with that view, the society as a whole has no such position. On the contrary, it is discussing how basic science can be protected from the influence of military funds. We apologize.

The News story 'Mysteries of Indian monsoon probed' (*Nature* **558**, 493–494; 2018) gave the incorrect name for the 5-year study. MISO-BOB was the name of just one component, not the whole study.



Bat man

Neuroscientist Nachum Ulanovsky uses fruit bats and a long, dark tunnel to study how the brain navigates.

BY ALISON ABBOTT

On a sun-parched patch of land in Rehovot, Israel, two neuroscientists peer into the darkness of a 200-metre-long tunnel of their own design. The fabric panels of the snaking structure shimmer in the heat, while, inside, a study subject is navigating its dim length. Finally, out of the blackness bursts a bat, which executes a mid-air backflip to land upside down, hanging at the tunnel's entrance.

Nachum Ulanovsky, the study leader, looks affectionately at the creature as his graduate student offers it a piece of banana — a reward for the valuable data it has just added to their

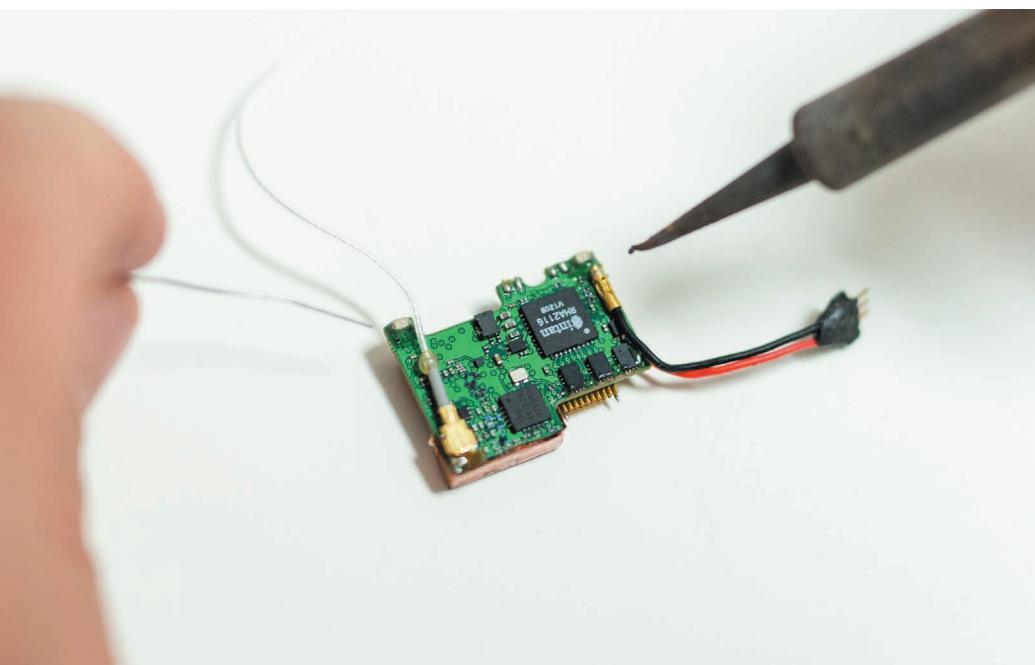
latest study of how brains navigate.

The vast majority of experiments probing navigation in the brain have been done in the confines of labs, using earthbound rats and mice. Ulanovsky broke with the convention. He constructed the flight tunnel on a disused plot on the grounds of the Weizmann Institute of Science — the first of several planned arenas — because he wanted to find out how a mammalian brain navigates a more natural environment. In particular, he wanted to know how brains deal with a third dimension.

The tunnel, which Ulanovsky built in 2016, has already proved its scientific value. So have the bats. They have helped Ulanovsky

DAVID VAKNIN FOR NATURE

Nachum Ulanovsky
with one of his
research bats.



A neural logger designed for wireless recording of neurons in flying bats.

to discover new aspects of the complex encoding of navigation — a fundamental brain function essential for survival. He has found a new cell type responsible for the bats' 3D compass, and other cells that keep track of where other bats are in the environment. It is a hot area of study — navigation researchers won the 2014 Nobel Prize in Physiology or Medicine and the field is an increasingly prominent fixture at every big neuroscience conference.

“Nachum’s boldness is impressive,” says Edvard Moser of the Kavli Institute for Systems Neuroscience in Trondheim, Norway, one of the 2014 Nobel laureates. “And it’s paid off — his approach is allowing important new questions to be addressed.”

And for brain scientists hitting the limits of what they can learn from highly simplified behaviour in the lab, Ulanovsky is a pioneer of ‘natural neuroscience’. Over the years, his arenas and tunnels have been getting larger, more sophisticated and less like an artificial lab environment. Up next is a giant maze that will allow his team to ask even more advanced questions about how the brain copes with making decisions — such as which way to turn — on the wing. “If we want to really understand how the brain works, we need to study animals doing more natural tasks,” says Dora Angelaki, a neuroscientist at Baylor College of Medicine in Houston, Texas. “More of us are finally starting to realize this.”

ARMED FOR SCIENCE

When Ulanovsky opened his lab at the Weizmann Institute in 2007, he was completing a circular flight path of his own. His family emigrated from Moscow to Israel in 1973, when he was just four months old, and

settled in Rehovot. As a child, Ulanovsky played in the Weizmann’s subtropical gardens and attended science events for local children and young people.

Once they turn 18, most physically fit Israelis enter compulsory military service. But Ulanovsky didn’t want to lose academic momentum when he graduated from high school at 16, so he enrolled in a three-year physics course at Tel Aviv University — even though that meant starting his military service late and, as a result, serving for a longer period.

His service proved productive. In addition to getting general military training, he was put in a research and development division because of his physics background. Over five years, he learnt technical skills such as designing high-tech instruments and programming that would later prove invaluable in designing arenas and sensors for his bats. The army allowed him time off to take courses that supported his growing interest in biology. He left the army intent on becoming a neuroscientist, and launched into a PhD at the Hebrew University in Jerusalem, studying how the cat brain processes auditory signals.

He discovered that auditory neurons have their own type of memory, and promptly immersed himself in the voluminous memory literature, where he discovered the overlapping field of navigation (animals have to remember where they have been to navigate, and it is not by chance that memory and navigation are processed in the same brain area). The field was dominated by studies in ground-based rats and mice, whose navigational experience is relatively easy to measure as they scuttle around small boxes in labs. But the question of how different

animals perceive the world as they move vertically — swimming, climbing trees or flying — had not been seriously addressed. Ulanovsky decided that to study the brain’s complex navigational code more holistically, he needed a mammal whose route-finding experience is mostly 3D, which led him to the only flying mammal: the bat.

He joined a bat lab at the University of Maryland in College Park to learn more about the creatures. He found several similarities to rodent models of navigation, discovering that bats, too, use special cells to get around¹. By 2007, Ulanovsky had his own bat lab and a tenure-track position at the Weizmann.

Ulanovsky is a composed person, but his equanimity can wobble when he talks about bats. His voice gets louder by a few decibels, and his face lights up. “In the West, people are frightened by creatures of the night — in Hollywood movies, when the heroine goes into a dark building and bats come rushing out, you know something bad is going to happen.” The fear is misplaced, he says. “In China, bats are considered a good omen.”

SPACE ODYSSEY

Neuroscientists have been mesmerized by how the brain encodes its spatial environment ever since the 1970s, when John O’Keefe at University College London found that the rat brain had a neat way to know where the animal is². When he placed electrodes in a region of the brain called the hippocampus, O’Keefe found neurons that fired only when a rat was in a particular location in its enclosure, creating a sort of cognitive map. He called them ‘place cells’.

Nearly three decades later, Edvard Moser and May-Britt Moser, also at the Kavli Institute, discovered another type of wayfinding cell in the nearby entorhinal cortex: grid cells, which fire not just at a single place in the enclosure, but at multiple points arranged in a hexagon³. These cells make up a brain code that allows the animal to keep track of its relative position in space, much like a tiny Global Positioning System (GPS). The Mosers shared the 2014 Nobel prize with O’Keefe; they and other scientists have also discovered other types of navigation cell in the hippocampal area, including those that fire in response to head direction⁴, or to a border such as a cage wall⁵.

Almost all of these discoveries came from rats: animals that — aside from, say, raising themselves on their hind legs to sniff, or accidentally falling from shelves — live their lives on the horizontal. One imaginative attempt to get around this monitored rats with implanted electrodes in weightless conditions during a 1998 flight on a NASA space-shuttle, but the result was inconclusive⁶.

For Ulanovsky, the virtues of bats extended beyond the animals’ suitability for understanding 3D mapping: he wanted

to work with a wild animal, to build a better picture of natural behaviour. He started to think that highly controlled lab experiments, so crucial to understanding some basic properties of neurons, needed a reality check. “We don’t know nearly enough about how all these cells work together to map the environment that animals inhabit in the wild,” he says. So he reasoned that bats caught from the wild and flown in less constrained environments would be the ideal subjects. Moreover, Ulanovsky was convinced that studying the system in something other than a lab rodent would help to identify which aspects of behaviour cut across species.

Edvard Moser agrees that studying the same skill in many species is important. “Knowing the different ways it is possible to solve the same problem will help us learn in general terms how brains, including the human brain, work.”

BAT CAVE

Before Ulanovsky could put his ideas to the test, he had to find the right sort of bat, check how it explored its natural environment and, most challengingly, design instruments to collect data from the bat and its brain.

Data from the brains of rats running around small enclosures are generally picked up by implanted electrodes and transferred to computers using cables. “Clearly, that won’t work in flying bats,” says Ulanovsky. He set about designing wireless GPS and electrophysiology devices that are small enough for a bat to carry. It was a technical challenge, and he might not have succeeded without his army training in instrumentation and software, he says.

His GPS logger is a 5-square-centimetre device tipping the scales at 8 grams. His neural logger, with 16 spindly electrodes — each thinner than a human hair — weighs in at just 7 grams. It is sensitive enough to record several individual neurons firing, and it can store many hours’ worth of data.

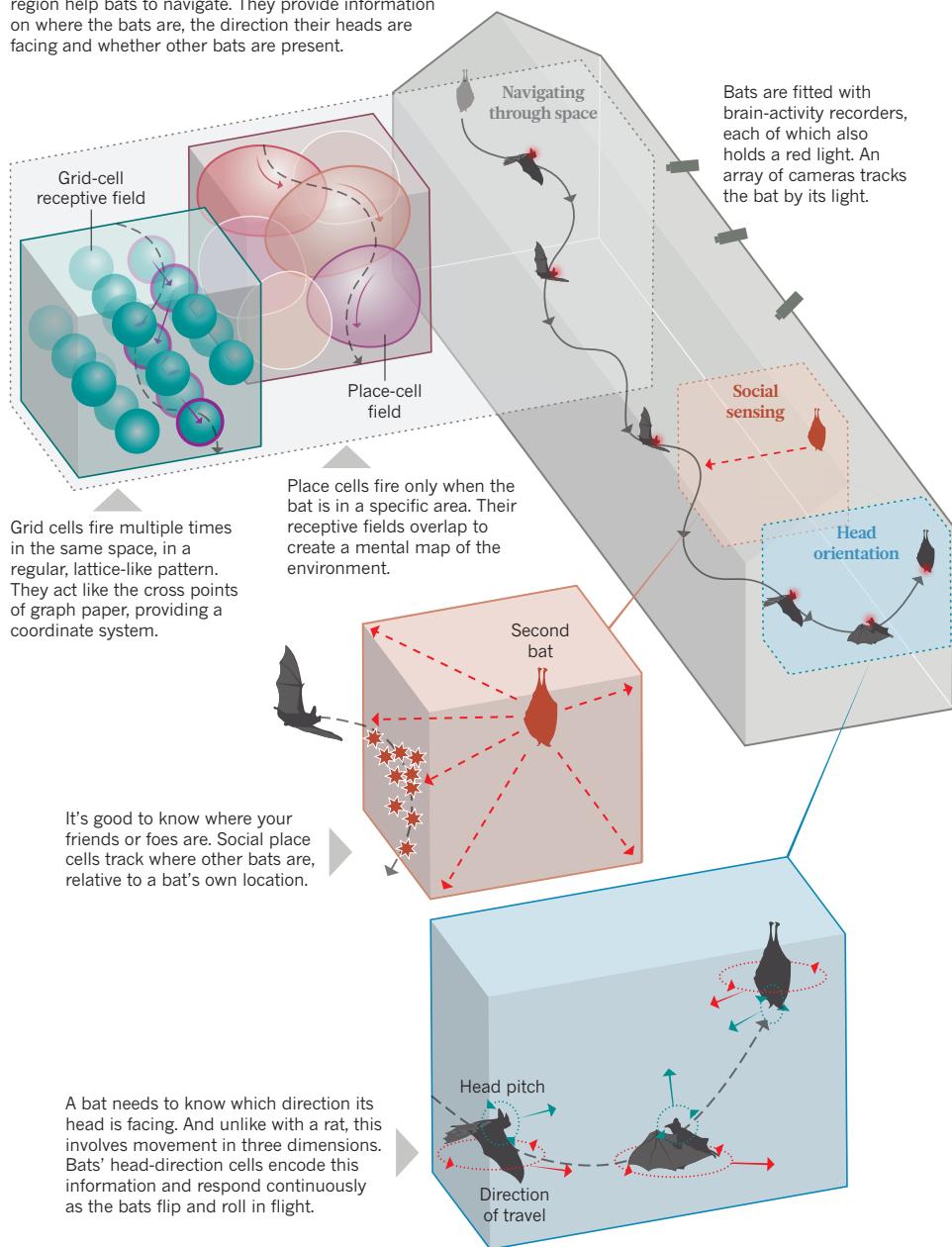
Tiny as they are, these loggers are too heavy for many bats to carry — including the delicate 20-gram bat *Eptesicus fuscus*, commonly known, ironically, as the big brown bat, and the species Ulanovsky studied when he was at Maryland. Instead, he settled on using the Egyptian fruit bat (*Rousettus aegyptiacus*). It’s ten times larger, approaching the size of an average laboratory rat, and common in Israel. “That was the low-tech part of my approach to miniaturization — choose a bigger bat,” says Ulanovsky.

Some bats can be vicious, but Egyptian fruit bats, he says, “are easy to tame and very nice to work with”. A couple of times a year, he picks up a giant net and heads out on a bat-catching safari, collecting specimens from colonies that inhabit abandoned buildings, or caves in the Judean hills.

One of his earliest experiments, started

Flight trackers

Several groups of cells in the hippocampal brain region help bats to navigate. They provide information on where the bats are, the direction their heads are facing and whether other bats are present.



in 2008, aimed to find out how far his bats chose to fly when left to their own devices. Very little was known about the natural behaviour of bats, he says, so he needed to gather some basic information. He armed 35 bats with GPS loggers and discovered that they flew 15 kilometres or more each night to find dinner — remembering the exact location of a particular heavily fruited tree⁷.

He also built flight rooms in his labs (see ‘Flight trackers’). The largest is about 6 × 5 × 3 metres — close to half the size of a squash court — and is decked out with cameras, landing balls for the bats to hang from and feeding stations where they can be tempted with fruit. Clad in metal and a layer of black acoustic foam to shield it from

external noise and electrical signals, the room is silent. The lighting can be adjusted from dim to very dim.

In the control room next door, the bats appear as tiny dots of light moving across a screen. Each bat carries a red light-emitting diode (LED), tracked by the cameras as the animals flit about the room. Their brain activity is monitored with a neural logger whose electrodes are surgically implanted into the hippocampus and whose external hardware is fixed to the skull with tiny screws. The cameras and loggers enable Ulanovsky to correlate the firing of neurons with the bats’ exact position in space.

In this set-up, he has been able to reveal the 3D territory of a typical bat-nav neuron. For

example, place-cell fields — measured in rats as flat circles of a particular size — turned out in flying bats to be almost spherical⁸, showing none of the vertical elongation that some rat experiments had predicted⁹. He worked out how head-direction cells operate as a 3D compass¹⁰, and discovered another type of navigation cell — the long-sought vector cell — which tracks angle and distance to a particular goal¹¹. One series of experiments helped put to rest a once-popular theory from rat studies that proposed that a certain type of brain oscillation creates grid-like neural maps; the oscillation turned out to be absent in bats, and therefore not necessary for such map-building¹².

He also explored the influence of a bat's social world. When he put a companion bat into the flight room, he discovered that

fruit trees. But Ulanovsky's team has tried to recreate some of the features that the brain uses as navigational aids. Graduate student Tamir Eliav collected a variety of objects and scattered them at intervals along the tunnel for the bats to use as fixed points in their internal map. Walking along the tunnel's length in the low glow of a dim LED strip light, past an old chest of drawers and a rusting bicycle rack, feels like being in an art installation.

Since the inaugural flight in March 2016, Ulanovsky and his students have collected data from more than 200 neurons across different bats. These early data hint at interesting insights. For example, Ulanovsky found that a single cell would fire at one location in a small area but also at a quite different location in a large area, indicating that place cells might represent multiple spatial scales, not just one particular scale. Researchers hadn't been able to spot this pattern in experiments in small enclosures. Ulanovsky needs more data to confirm this, but it would be in line with the predictions of some theoreticians. "If place cells all had small, laboratory-sized place fields, there would not be enough neurons in the hippocampal area to individually cover the great distances that bats travel," says Ulanovsky, "so it makes sense that some place cells respond to multiple scales."

TUNNEL VISION

That's motivated him to design a bigger and better tunnel. Earlier this year, a private sponsor provided half of the 9 million shekels needed to build a kilometre-long tunnel with more densely positioned, wired antennas. This will allow measurement of even larger place fields, with more precise 3D localization. This tunnel will have a 15-metre side branch to allow the scientists to study how the same neurons respond to short and long flights, and how the brain stitches together these two scales. Air conditioning will allow experiments to run throughout the blistering summer.

The tunnel and its once-wild bats represent a useful halfway house between the real world and the lab, says Angelaki, who researches spatial navigation and decision-making in the brains of mice and monkeys.

"Behavioural neuroscientists like myself are increasingly realizing how important it is to move away from overtrained lab-animal brains," she says. In typical lab experiments, animals are trained in a very specific, usually unnatural, task. "That may not have anything to do with how that animal has evolved brain connectivity to optimize foraging in the wild," she says.

Like others around the world, Angelaki's lab is starting to use neural loggers to monitor more natural rodent behaviour, such as foraging for food scattered in their enclosures. She predicts that more researchers will start setting up their experiments with an eye on the natural world. "Over the next five years or so, results will start to emerge and there will be a big change in neuroscience practice," she says.

However, as Moser notes, Ulanovsky's bats aren't yet doing anything as clever as finding a fruit tree in the wild. "It doesn't take much thought to fly up and down a tunnel," he says. So Ulanovsky is nursing an even bigger mind-reading ambition. He is seeking funding for a maze 40 metres wide and 60 long — a little under half the size of a football pitch — to test how bat brains represent more complex environments, then plan and make decisions about how to navigate them.

The maze will be made up of interconnected tunnels in which the bat won't always be able to see its goal (usually a food treat such as a piece of banana). It will instead have to rely on memory in its cognitive map. Ulanovsky has a series of increasingly complex experiments in mind — setting up multiple goals, for example, or suddenly blocking a path that the bat had memorized. He has questions about how bats choose between several goals, or recompute a path, or how cells respond when a bat loses its way. "Do the vectors in the brain start rotating wildly?" he wonders. "These are all fascinating questions to which we have no answers."

And the bats are obliging subjects. On a good day in the tunnel, a bat can soar and wheel for thousands of metres before taking a break for its banana. "They are misunderstood creatures," says Ulanovsky, standing at the end of the tunnel and gazing at a just-landed bat with obvious tenderness. "And they will help science." ■

Alison Abbott is *Nature's* senior European correspondent.

1. Ulanovsky, N. & Moss, C. F. *Nature Neurosci.* **10**, 224–233 (2007).
2. O'Keefe, J. & Burgess, N. *Brain Res.* **34**, 171–175 (1971).
3. Hafting, T., Fyhn, M., Molden, S., Burgess, N. & Moser, M.-B. & E. I. *Nature* **436**, 801–806 (2005).
4. Taube, J. S., Müller, R. U. & Ranck, J. B. Jr. *J. Neurosci.* **10**, 420–435 (1990).
5. Solstad, T., Boccara, C. N., Kropff, E., Burgess, N. & Burgess, E. I. *Science* **322**, 1865–1868 (2008).
6. Knierim, J. J., McNaughton, B. L. & Poe, G. R. *Nature Neurosci.* **3**, 209–210 (2000).
7. Tsoar, A. *et al.* *Proc. Natl. Acad. Sci. USA* **108**, e718–724 (2011).
8. Yartsev, M. M. & Ulanovsky, N. *Science* **340**, 367–372 (2013).
9. Hayman, R., Verriots, M. A., Jovalekic, A., Fenton, A. A. & Jeffery, K. J. *Nature Neurosci.* **14**, 1182–1188 (2011).
10. Finkelstein, A. *et al.* *Nature* **517**, 159–164 (2015).
11. Sarel, A., Finkelstein, A., Las, L. & Ulanovsky, N. *Science* **355**, 176–180 (2017).
12. Yartsev, M. M., Witter, M. P. & Ulanovsky, N. *Nature* **479**, 103–107 (2011).
13. Omer, D. B., Maimon, S. R., Las, L. & Ulanovsky, N. *Science* **359**, 218–224 (2018).

"We don't know nearly enough about how all these cells work together to map the environment."

the monitored bat had 'social place cells' that track the companion's position¹³. He'd imagined that such cells must exist somewhere in the brain — bats obviously need to know where their fellow bats are, as well as their predators — but was not expecting they would necessarily show up inside the hippocampus. He is now monitoring how the brains of two or three bats register the social interaction of up to ten companion bats living together in the large flight room for several months.

But Ulanovsky's burning question was how this set of navigation cells would perform outside a flight room, during more natural behaviour. It would be impossible to monitor the positions of bats in the wild — cameras would be no use because the bats' ranges are too large, and GPS would not give high enough resolution — so Ulanovsky decided that an artificial tunnel was the best option.

As a bat flies through the 200-metre tunnel, he can monitor its exact position using a tiny signalling device on the bat itself and a suite of 15 antennas placed at intervals outside the structure to pick up its radio transmissions. Each antenna sends its computed distance from the signalling tag by Wi-Fi to a workstation at the tunnel entrance, where the full 3D movement of the bats is recreated. The whole set-up cost around 900,000 Israeli shekels (US\$250,000) to construct.

From the bats' point of view, flapping through the tunnel is much easier than a 15-kilometre night-time foray to distant

COMMENT

CITIES Cape Town ran dry because of politics, not climate change **p.174**



TOXICITY How citizen science brought justice to the people of Flint **p.180**

DEVELOPMENT Ghana leads the way in home grown diagnostics kits **p.181**

CHEMISTRY Happy 200th to hydrogen peroxide, industrial workhorse **p.181**

ADAM DEAN/THE NEW YORK TIMES/REDUX/EYEVINE



A Rohingya refugee from Myanmar in a Bangladeshi camp in 2017.

Grand challenges for humanitarian aid

Fund and study these priorities for natural and social sciences to meet a gaping need, urge **Abdallah S. Daar, Trillium Chang, Angela Salomon and Peter A. Singer**.

The gap between the magnitude of humanitarian need and the global capacity to respond is massive and growing. Here we describe an attempt to map ways in which that gap might be closed (see 'Top 10 Humanitarian Grand Challenges').

Humanitarian crises directly affect more than 140 million people in 37 countries, according to the United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA). More than 65 million of these people have been forcibly displaced from their homes — the

highest level since the Second World War. Nearly 60% are currently in Africa and the Middle East, including in Turkey, Lebanon, Uganda and Ethiopia¹. The rest include refugees, asylum seekers, people displaced internally, those not yet seeking asylum and many more.

Much of this humanitarian need derives from violent conflicts and civil wars that target civilians and their support systems, including shelters and hospitals. Much also follows natural disasters such as earthquakes, hurricanes, floods and drought. With climate change, it is highly likely that

some of these disasters will get worse and more frequent¹.

All of these people need aid, and the funds available are increasingly inadequate¹. Just one-third of the US\$25.4 billion required for humanitarian aid for 2018 will be covered. In other words, the current humanitarian system is buckling. It desperately needs much more programme funding to close the gap. At the same time, it needs more funding for innovative solutions: uses of technology, products and processes from other sectors; new forms of partnership; and drawing on the ideas and coping capacities of ▶

Top 10 Humanitarian Grand Challenges

RANK	PRIORITY IDENTIFIED	RESEARCH QUESTIONS
1	<h2>Strengthen economies</h2>	<p>Restore functioning markets and the economic stability of affected communities by:</p> <ul style="list-style-type: none"> Scaling up cash-based assistance (rather than in-kind commodities) Improving access to financial services Increasing autonomous choice over spending Expanding social safety-net programmes, such as provision of health care, shelter and transport Engaging cross-border refugees, particularly women, who are displaced to countries where they are forbidden to work outside camps*
(147 cumulative score)	<h2>Reduce inequality</h2>	<p>Strengthen resilience in communities at risk of humanitarian crises by:</p> <ul style="list-style-type: none"> Reducing inequality and poverty Promoting gender equality Improving education*
(141)	<h2>Improve metrics</h2>	<p>Measure effectiveness of humanitarian aid by moving away from metrics that measure 'cost-per-beneficiary' to those that measure how the needs are met of:</p> <ul style="list-style-type: none"> the most vulnerable the most systematically excluded the hardest-to-reach communities
(138.5)	<h2>Address funding</h2>	<ul style="list-style-type: none"> Shift from short-term emergency funding toward longer-term humanitarian financing Ensure accountable, impactful investments that include incentives or subsidies for host governments to contribute alongside foreign assistance*
(128)	<h2>Protect identity</h2>	<p>Provide affected persons with an official private, secure digital identity that reduces the risk of creating stateless persons.</p> <p>This might:</p> <ul style="list-style-type: none"> Incorporate a universal health card Safely and privately store, transport, validate authenticity of, and disseminate personal documents (such as bank cards, land deeds, birth certificates, school diplomas and medical records)*
(121)		

RANK	PRIORITY IDENTIFIED	RESEARCH QUESTIONS
6	Expedite aid	<ul style="list-style-type: none"> What are the most effective international mechanisms and auspices under which to engage governments to develop partnerships for immediate disaster/emergency relief? How feasible and effective are crowdfunding platforms to speed the availability of money in crisis situations? How can mechanisms for regional neutral bodies to intervene rapidly in the case of disasters be better coordinated? How can the voices of those affected by crises be amplified most effectively?
(119)		
7	Save more lives	<ul style="list-style-type: none"> What methods promote and ensure compliance (of non-governmental organizations, governments) with international humanitarian law? How can such laws be strengthened? How can the private sector improve the delivery of aid and increase the speed, effectiveness and cost-efficiency of delivering or manufacturing commodities (such as by 3D printing) in hard-to-reach places? How can crisis-affected people be supported or empowered to create their own local solutions — such as by locally manufacturing and reusing items? In what ways can military know-how and capabilities, including transport and logistics, be used ethically in disaster responses? What are potential political obstacles, and how can they be overcome?
(117)		
8	Support mental health	<ul style="list-style-type: none"> How effective are culturally sensitive and locally applicable emergency intervention programmes based on the World Health Organization's Mental Health Gap Action Programme for mental health and psychosocial support? Where are there gaps and how can they be filled? What are the most effective ways for health-care providers to advocate for the incorporation of established ethical principles and more counselling into emergency mental-health intervention programmes? What are the population metrics and outcome indicators for mental-health policy and programme surveillance? Can artificial intelligence (such as chatbots or apps) deliver mental-health and psychosocial support, in a culturally sensitive and effective manner?
(116)		
9	Democratize data access	<ul style="list-style-type: none"> What culturally specific and community-based strategies will efficiently and effectively integrate crisis-affected people with worldwide data sources? How can mobile-network operators become valuable contributors to preparedness before, and responses after, humanitarian disasters? How effective are existing innovative ways to share data in humanitarian settings, such as mesh networks, bluetooth technology, microwave technology and peer-to-peer networks? What other novel strategies exist?
(113)		
10	Boost direct communication	<ul style="list-style-type: none"> What are examples of low-cost satellite or other technologies that can facilitate logistics and cut response time in crisis settings, and how effective are they? How can non-governmental organizations, governments and other actors gain feedback from affected persons to improve humanitarian responses? How effective are online surveys, feedback apps and chatbots? What other novel solutions exist?
(110.5)		

*Challenge reformatted and/or slightly reworded from the original submission to increase clarity and coherence.

START AT A SUMMIT

What are Global Alliance for Humanitarian Innovation and Grand Challenges Canada?

The need for innovation in the humanitarian space was recognized at the World Humanitarian Summit in Istanbul¹⁰ in May 2016. The largest ever United Nations gathering, this had 9,000 participants from at least 173 countries, including 55 heads of state and governments, hundreds of private-sector representatives, and thousands of people from civil society and non-governmental organizations, including multilateral development banks such as the World Bank.

The summit created the Global Alliance for Humanitarian Innovation with the mission of achieving higher impact and efficiency in humanitarian action¹¹. It complements several initiatives, including Global Humanitarian Lab, Global Partnerships for Humanitarian Impact and Innovation, and the Canadian Humanitarian Assistance Fund. Unfortunately, many of these

have insufficient funding to address the magnitude of the problem by creating a healthy pipeline of seed innovations; most do not have the capacity to scale them up.

Grand Challenges Canada (GCC), supported by the Government of Canada, funds technological, social and business innovations in global health. Since its founding in 2010, GCC has supported 1,000 projects in more than 80 countries (see go.nature.com/2jyaobz). The leaders of GCC have a track record of partnering to identify priorities that catalyse the creation of impactful research funding programmes at the global level. These include: the Bill & Melinda Gates Grand Challenges in Global Health programme, based on a 2003 study⁵; the Global Alliance for Chronic Diseases, based on a 2007 study⁶; and the Global Mental Health Initiative of the US National Institute for Mental Health and GCC, based on a 2011 study⁷. **A.S.D., T.C., A.S. & P.A.S.**

► crisis-affected people — in a way that is iterative and rigorously evaluated^{2,3}. A balance of the two types of funding would help the humanitarian system to become more efficient and more effective.

To this end, humanitarian agencies met with governments, private-sector representatives, philanthropists and affected persons (people who were refugees, were born in refugee camps or who worked closely with such affected people) in November 2016 in Toronto, at an event convened by Grand Challenges Canada (GCC; see 'What are Global Alliance for Humanitarian Innovation and Grand Challenges Canada?'). Participants agreed that innovation in the humanitarian sector would be catalysed by a list of priorities, systematically identified and agreed upon. Here we set these out, and describe how they were reached.

LAYING FOUNDATIONS

Participants agreed on the definition of a grand challenge as a specific critical barrier that, if removed, would help to solve an important humanitarian problem. They agreed that humanitarian assistance is aid and action designed to save lives, alleviate suffering and maintain and protect human dignity during, and in the aftermath of, human-made crises and natural disasters, as well as aid and action to prevent such situations or prepare for them. Participants advocated the empowerment of affected communities, especially women and girls,

and the inclusion of actors beyond the usual humanitarian community — such as youth, the private sector and affected persons. Finally, participants specified that action should be governed by the four humanitarian principles — humanity, impartiality, neutrality and independence (see go.nature.com/2kb88h7).

The participants also foresaw the need to create partnerships around the identified priorities. The Toronto launch meeting culminated with a 'dry run' to identify a few potential grand challenges (see go.nature.com/2tjms5k). The participants asked GCC to identify grand challenges in humanitarian innovation through the Delphi method⁴, a technique that builds consensus

“Identifying priorities, as in this study, is just the first step.” using iterative feedback from dispersed experts. The participants also agreed to serve as the nucleus of the Delphi panel. There followed a three-round Delphi study, similar to previous exercises in health and disease⁵⁻⁷.

GCC built a panel of 68 experts in humanitarianism and innovation. Those who had taken part in the Toronto launch suggested one or two names from their networks to join (a technique called snowball sampling). Also invited were ten affected individuals who had attended the World Humanitarian Summit in Istanbul, Turkey, in May 2016 and were

recommended for the Delphi panel by an official of UNOCHA. These included, for example, the founder of the Feminist Dalit Organization, which represents the discriminated group in Nepal known as 'untouchables'; a refugee ambassador for the Office of the United Nations High Commissioner for Refugees (UNHCR); and a nominee for the 2015 US Secretary of State's International Women of Courage award.

Sixty people took part in at least one of the three rounds. The first launched in January 2017; the last closed in July 2017. All communication with panellists was through e-mail.

In the first round, each participant answered the following question: 'What one grand challenge, if solved, would make humanitarian work more effective and efficient for the long term?'

Panellists submitted 106 answers. We lightly edited these to ensure consistency (deleting duplicates and collating analogues). This generated 83 unique statements. These we grouped into categories including: financing, economic empowerment of affected communities, gender equity/gender-based violence, digital identity, documentation/data management and tools.

In round two, panellists chose 20 challenges of these 83, and ranked them from 20 (highest priority) to 1 (lowest). Scores for each statement were then summed across all participants to identify an overall top 20.

In round three, the panellists ranked their top 10 from these 20, with 10 being the most important. Scores for each statement were then summed across all participants to identify an overall top ten list (see table, in which the rankings have been inverted so that the priority with the highest cumulative score is ranked first). In the second and third rounds, participants were encouraged to add comments or suggestions for rewording or combining statements.

Of the 60 panellists who participated in one or more rounds, 50 completed round three (83.3%). This is a high response rate for this type of large-scale international study with dispersed participants. Only four of the ten invited affected persons participated in all three rounds of the study; the others struggled to take part because they were still living and working in crisis situations (a difficulty not unique to this study). For the other participants, the distribution of organizations, geographical regions, gender and expertise was not significantly different between those who completed all three rounds and those who did not.

For the final ten Humanitarian Grand Challenges, participants were also asked to suggest potential research questions (see



People stand behind a safety cordon in San Juan Aotenango, Guatemala, after the nearby Fuego volcano erupted in June.

table). Some challenges — denoted with an asterisk (*) — were reformatted and/or slightly reworded from the original submission to increase clarity and cohesiveness. Many participants also suggested important practical steps that are more actions than research questions (see Supplementary Information).

In a conference call in September 2017, panellists discussed designing and implementing partnerships and large-scale innovation funding programmes to address the research questions.

NOW WHAT?

The ten humanitarian grand challenges identified in this study encompass many sectors. They call for a reduction of the distinction between humanitarian and development efforts. Of course, these do not cover all potential barriers or gaps in either realm. Some of these challenges are long-standing. To be addressed, they now need collaborative thinking, large-scale funding and leveraging of new technologies. Some have been tackled in ways that could do with more impact evaluation — for example, identifying the best way to deliver supplementary nutrition for poor families with young children in food crisis situations².

Identifying priorities, as in this study, is just the first step in a long-term, continuous process of trying to effect change. Ideally, most of the research questions provided here will examine how to scale up solutions in specific locations, and in ways that will inform the global humanitarian community and help it to prepare for future emergencies.

Already, this initiative has begun to make a difference. In February this year, the US Agency for International Development's Office of US Foreign Disaster Assistance (USAID OFDA), the UK Department for International Development (DFID) and Grand Challenges Canada launched a multi-million-dollar initiative to support innovations that engage the private sector and involve affected communities to provide, supply or locally generate safe water and sanitation, energy, life-saving information, or health supplies and services to help conflict-affected people. This initiative is called Creating Hope in Conflict: A Humanitarian Grand Challenge (see go.nature.com/2kscfa2).

Within 2 months, 615 proposals were received from 87 countries; approximately 300 came from low- and middle-income countries, including more than 100 from countries in active conflict. Seed grants of around US\$250,000 will be awarded to pilot projects, and grants of up to \$1 million will be awarded to a select number of 'transition-to-scale' projects. Announcements are expected in late 2018. More funding partners are likely to join soon.

Once a robust pipeline of innovations is established, the challenge will be to scale them in a sustainable manner. With sufficient funding and effective partnerships, we hope to see progress on the priorities identified here. ■

Abdallah S. Daar is professor of clinical public health, global health and surgery at

the University of Toronto, Canada; chair of the International Scientific Advisory Board of Grand Challenges Canada; and a permanent fellow of the Stellenbosch Institute for Advanced Study, South Africa. **Trillium Chang** is a master's student in public health at the University of Cambridge, UK. **Angela Salomon** is a master's student in public health at the University of Toronto. **Peter A. Singer** is chief executive of Grand Challenges Canada and professor of medicine at the University of Toronto and the University Health Network, Canada. e-mail: a.daar@utoronto.ca

1. UNHCR. *Global Trends: Forced Displacement in 2015* (UNHCR, 2016).
2. Aladysheva, A. 'Why humanitarian assistance needs rigorous evaluation' (Stockholm International Peace Research Institute, 2018); available at <https://go.nature.com/2talahf>
3. Betts, A. & Bloom, L. *Humanitarian Innovation: The State of the Art*. UNOCHA Policy and Study Series 009 (OCHA, 2014).
4. Linstone, H. A. & Turoff, M. (eds) *The Delphi Method: Techniques and Applications* (2002); available at <https://go.nature.com/2jeutgj>
5. Varmus, H. *et al.* *Science* **302**, 398–399 (2003).
6. Daar, A. S. *et al.* *Nature* **450**, 494–496 (2007).
7. Collins, P. Y. *et al.* *Nature* **475**, 27–30 (2011).
8. Tangcharoensathien, V., Thwin, A. A. & Patcharanarumol, W. *Bull. World Health Organ.* **95**, 146–151 (2017).
9. Khanna, T. & Raina, A. 'Aadhaar: India's 'Unique Identification' System' Harvard Business School Strategy Unit Case No. 712-412 (2012).
10. World Humanitarian Summit. *Commitments to Action* (WHS, 2016); available at <https://go.nature.com/2krir39>
11. Global Alliance for Humanitarian Innovation. *Stakeholder Consultation Report* (2017); available at <https://go.nature.com/2u6ndpe>

Supplementary information accompanies this article: see go.nature.com/2trwopy



RODGER BOSCH/AFP/GTET

The narrow body of water that remained at South Africa's Theewaterskloof Dam in May 2017.

Lessons from Cape Town's drought

Don't blame climate change. People and poor planning are behind most urban water shortages, argues **Mike Muller**.

Since May, winter rains have brought a reprieve to the citizens of Cape Town, South Africa. The city had endured severe drought for three years. Concerns that its water supply might run out in the summer have been set aside, hopefully, for another year. But the city remains vulnerable.

The situation was very different in 2013. Then, Cape Town had one of its highest annual rainfalls in decades. Reservoirs brimmed, and officials declared there was no need to increase supplies before the 2020s. After another wet winter in 2014, the 6 main reservoirs that feed the city were 97% full.

Then the drought began. Reservoir levels fell to 71% in 2015 and to 60% in 2016 (see 'Cape Town drought'). When they reached 38% in 2017, at the beginning of what

looked set to be a long, hot summer, people began to panic.

Municipal authorities told residents to slash their water consumption. For suburban households, that meant going from pre-drought usage of around 200 litres per person per day to 50 litres per person per day (picture a bathtub filled to less than 10 centimetres). Although many of their poorer compatriots regularly live with such a supply, suburbanites suddenly had to give up their gardens and collect shower water to flush their toilets. The city more than halved its overall use, to just over 500 million litres a day, and avoided 'day zero'.

Cape Town is one of several cities to see its water supply fail in the past decade. In 2014 and 2015, parts of São Paulo in Brazil

received water for only two days a week. Once the city's reservoirs had been drained of clean water, the utility firm pumped and treated the polluted water that remained. In 2008, Barcelona in Spain had to ship water in from Marseille, France. During its decade-long 'millennium drought' in the 2000s, Australia spent billions of dollars on desalination plants, most of which have not been used since.

It is important to learn from the experiences of Cape Town and elsewhere. Urban growth means that many more places will face similar challenges as they compete with surrounding regions for water. Big cities need to begin informed long-range planning and to focus on minimizing risks from current climate variability. Climate change adds

to the uncertainties. Shortages attributed to extreme weather or to global warming are still more often due to poor management. People's beliefs and behaviour are as much a part of the systems to be managed as are pipes, pumps and the environment.

SHORT-SIGHTED

Cape Town's problems are due in large part to a turn away from management based on science and risk assessment towards a more populist approach¹.

Since the 1980s, South Africa's major conurbations have used systems models to guide their water management². These models, run by the national government, are considered world-class. They map links between river basins, reservoirs and transmission channels and use historical hydrological data to predict probable stream flows. Those are then matched to projections of demand to assess how much storage is needed. The models support real-time operations of the water network as well as planning for development. Crucially, they allow planners to assess risks of supply failures to different categories of users and evaluate the effectiveness of responses such as restrictions.

For two decades, policymakers heeded the models. They guided managers, for example, on when and where to tap sources and build reservoirs to enable the Western Cape Water Supply System (WCWSS) to meet rising demand from urban and industrial growth.

But dam building stalled in the 2000s, when local environmentalists campaigned to switch the focus to water conservation and management of demand. Such opposition delayed the completion of the Berg River Dam by six years. Eventually finished in 2009, the dam helped to keep the taps running in Cape Town this summer.

Back in 2009, the models had already flagged a need to boost Cape Town's water supplies after 2015, but officials dismissed the recommendations. They were happy to delay big capital investments and spend the money elsewhere. They missed that the Cape's wine and fruit farmers (who are entitled to one-third of the region's water) were not drawing their full allocation during the rainy years and, like the city's gardeners, were using more in drier years.

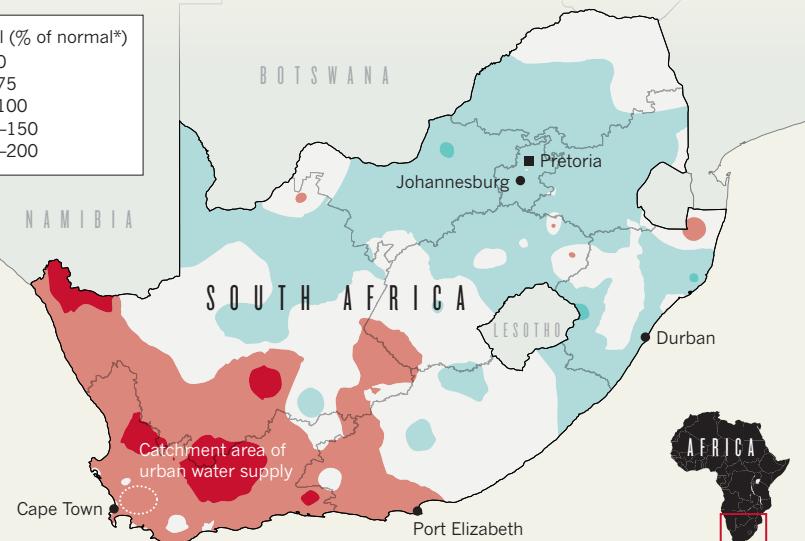
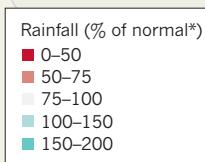
The response was short-sighted. The 6 Western Cape reservoirs that feed the city hold less than 2 years' supply: 890 million cubic metres, compared with a reliable annual yield of 570 million cubic metres. It took two successive dry winters, in 2015 and 2016, for the municipality to realize that it was in trouble. City leaders banned water use in gardens and car washing, and promoted conservation, water-efficient appliances and higher tariffs.

They defended their decisions. The councillor responsible for water services, Xanthea Limberg, wrote in April last year that they

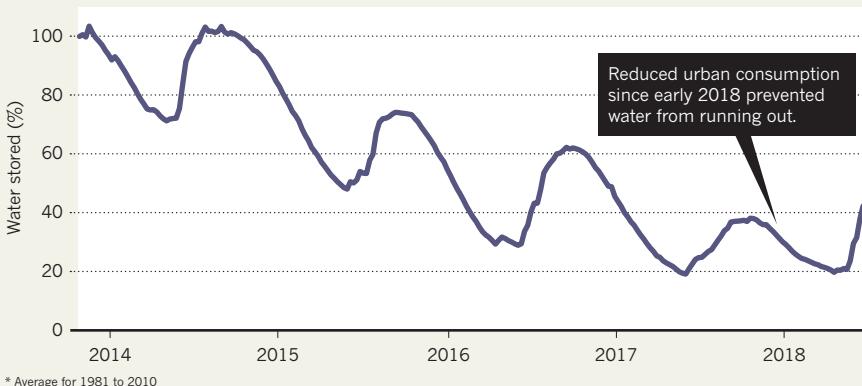
CAPE TOWN DROUGHT

Lower than average rainfall in South Africa's Western Cape exacerbated water shortages over the past 3 years. The total amount of water stored in the six largest reservoirs that supply Cape Town fell to new lows each year.

ANNUAL RAINFALL JULY 2016 TO JUNE 2017



REGIONAL RESERVOIR LEVELS



had not addressed the risk of such a severe dry spell because "it is not practical to ring-fence billions of rand for the possibility of a drought that might not come to pass".

Then the worst case happened — a very dry 2017. How dry is disputed. Rainfall and stream flows vary dramatically from place to place and year to year across the region's mountainous terrain. At the Jonkershoek weather station, which is close to the catchment of the two largest dams, rainfall varied: 1,250 millimetres in 2013; 900 millimetres in 2014; less than 500 millimetres in 2015; 750 millimetres in 2016; and 700 millimetres in 2017. In 2017, flows in a small, undeveloped stream in the same area were just 20% of what they were in 2013.

Three consecutive dry years have occurred before, in the late 1930s and from 1970. Three dry years in 2002, 2003 and 2005 were fortunately interrupted by a wet 2004. These risks were reflected in the hydrological models. But Cape Town's leaders did not comprehend the social and financial

implications of their decisions.

They do now. So far, direct costs of the water crisis — reduced water revenue, losses in agricultural jobs and production and indirect costs such as a drop in tourism — have come to more than 2.5 billion South African rands (US\$181 million). Water tariffs for consumers have been raised by 26% this year. Yet, investing 1 billion rands in infrastructure in 2013–14 would have cost just 75 million rands per year in interest charges — that would have been cheap insurance, even if it had proved unnecessary.

Cape Town's decision-makers have tried to shift the blame, with climate change an obvious target. Helen Zille, premier of the Western Cape, wrote last October that "the impact of climate change is probably the reason that climate cycles have become so unpredictable". Yet there is little evidence of a departure from normal variability in the catchments. Although data from outside sites are cited to support climate-change ▶



Residents fill water containers at the Newlands natural-water spring in Cape Town in November 2017.

WALDO SWIERS/BLOOMBERG/GETTY

► theories, it was the three-year sequence of dry years that proved devastating.

Yes, meteorological changes are expected over coming decades. For Cape Town, most climate models predict a decrease in rainfall by 2050, although local impacts are uncertain³. Stream flows are even less predictable, because they will be reduced by heat and aridity but can be increased by more-frequent and intense rainfall. Such trends and uncertainties can be factored into models, although historical hydrology should still provide reliable perspectives for the next few decades.

A more-immediate challenge for Cape Town is that the area that supplies the Western Cape water system is very small — less than 800 square kilometres. Local variations in climate, by themselves, call for a conservative and risk-averse approach, and the need for a diverse range of water sources to fall back on.

Now, Cape Town's leaders are working feverishly to build the schemes that were recommended back in 2009 for managing groundwater, reuse and surface supply. The crisis has obliged them, and others elsewhere in the country, to look more carefully at future challenges. The premier of Gauteng province, the country's inland urban hub, has convened a high-level task force to tell officials how to avoid Cape Town's experience.

GLOBAL PROBLEM

São Paulo and Barcelona also had precedents for their dry spells. And political decisions exacerbated their water crises. São Paulo's drought risk was highlighted in hydrological models, but wrangling between city, state and national governments delayed action for a decade⁴. In Barcelona, a surprise 2004 election win saw the Spanish Socialist Workers' Party (PSOE) stop a long-planned programme of dam development and river

transfers because of a manifesto commitment to regional allies⁵. In Australia, as in Cape Town, environmental opposition to dams and desalination increased cities' vulnerabilities to a multi-year drought.

China is a counter-example. It has managed to keep water flowing in some of the world's largest and fastest-growing cities through responsive government planning and major infrastructure projects. These include the Three Gorges Dam, which controls flooding on the Yangtze river, and the South–North transfer, which has channelled water from the Yangtze to Beijing since 2015. The nation has drawn some of those lessons from South Africa.

In 2002, Wang Shucheng, the Chinese minister charged with resuscitating the transfer project, visited the Lesotho Highlands Water Project. This binational network of tunnels and dams diverts water from the mountains of Lesotho to South Africa's inland economic hub. Back in China, Wang got the sequence right. He completed the major engineering projects needed to underpin supply while, in parallel, starting longer-term programmes to reduce pollution, manage demand and promote efficient water use⁶.

POLITICAL PROCESS

The greatest challenge for managers of urban water supplies is often getting political decisions made in a timely fashion, and with public support.

There is no universal best-practice approach to achieve this. Beyond

implementing strong centralized systems such as China's, improving cooperation between the various organizations involved might help⁷. Because rivers generally cross political boundaries, water management is often organized in 'watersheds' that can be distant from politicians and their citizens. Cape Town draws water from two rivers beyond its boundaries, each of which is managed by a different agency. Managing water in 'problem-sheds' that encompass major water users and the geographical areas on which they depend would be a better approach⁸.

As water needs grow and water systems evolve, more resources will need to be devoted to monitoring and modelling. Technical guidance must be integrated into political processes. As a minimum, politicians need to know who is doing the modelling and what the recommendations are.

They also need that information in a format and language that empowers them to act appropriately. So hydrologists must collaborate with experts from the social sciences and humanities, notably economics, policy and law, to develop water-management tools that decision-makers and the public can understand and use⁹. With greater involvement of other disciplines, it will be easier to ensure that appropriate social, economic and environmental criteria are used when selecting technical options — to store water in a dam or to use energy for desalination, for example.

Finally, practitioners also need to monitor and model peoples' behaviour. The long time scales over which decisions and interventions need to happen must also be better understood. And the easy resort to simplistic solutions to 'use less water' and 'rely on natural infrastructure' must be resisted.

Cities must move from crisis responses to effective management of the water that is essential to lives, livelihoods and environments. 'Day Zeros' are not inevitable. ■

As water needs grow and water systems evolve, more resources will need to be devoted to monitoring and modelling.

Mike Muller is an engineer and visiting adjunct professor at the Wits School of Governance at the University of the Witwatersrand in Johannesburg, South Africa.

e-mail: mikemuller1949@gmail.com

1. Muller, M. Civ. Eng. **June**, 11–16 (2017).
2. Basson, M. S. & Van Rooyen, J. A. J. Hydrol. **241**, 53–61 (2001).
3. Abiodun, B. J. *et al.* Clim. Change **143**, 399–413 (2017).
4. Escobar, H. Science **347**, 812 (2015).
5. Hernández-Mora, N., del Moral, L., La-Roca, F., La Calle, A. & Schmidt, G. Interbasin water transfers in Spain: Interregional conflicts and governance responses. In *Globalized Water* pp. 175–194 (Springer, 2014).
6. Wang, S. Resource-oriented water management: Towards harmonious coexistence between Man and Nature (China WaterPower, 2002).
7. Woodhouse, P. & Muller, M. World Dev. **92**, 225–241 (2017).
8. Mollinga, P., Meinen-Dick, R. S. & Merrey, D. J. Dev. Pol. Rev. **25**, 699–719 (2007).
9. Faticchi, S. *et al.* J. Hydrol. **537**, 45–60 (2016).



Adalbert Stifter's depiction of the west Hungarian floodplains of the Danube, painted around 1841.

CLIMATE SCIENCES

Imperial roots of climatology

Mott Greene applauds a history of how the Austro-Hungarian Empire shaped the field.

A world power ruled by the eccentric Habsburg monarchy, the Austro-Hungarian Empire was a force to reckon with for the frenetic 50 years preceding 1918. In the comprehensive, deeply researched *Climate in Motion*, historian of science Deborah Coen explores a lesser-known side of this unwieldy empire: its role as a crucible of modern climatology. Imperial scientists were a starry league, from meteorologist Julius Hann, who explored the relationships between prevailing wind, rainfall and mean temperatures, to geographer Alexander Supan, who established the global classification system for climate zones.

Coen probes imperial society and culture

to understand why the Austro-Hungarian scientific establishment and government devoted such vast resources to meteorology and climate sciences. She finds the driver in the political and social need to shore up an uneasy multinational alliance that incorporated what are now the Czech Republic and Slovakia, as well as parts of Poland, Italy and Romania, among other states. Historians are fond of saying that science is embedded in the context of a specific time and place. Coen demonstrates this unequivocally.

The sweeping narrative spans three-quarters of a century, from about 1850 to 1925. Coen has mined many archives to trace the intersecting careers of more

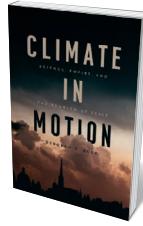
than a dozen major figures — meteorologists, botanists, geographers, geologists, painters and writers. Anton Kerner, for instance, saw how the geographical distribution of plants could help to determine climate history. Here, too, are Adalbert Stifter, an immensely popular novelist, travel writer and landscape artist; Emanuel Purkyně, Czech pioneer of microclimatology; and Wilhelm Schmidt and Felix Exner, among the first to conduct lab simulations of atmospheric circulation and its modification by topography.

Two themes knit together what is essentially a collection of interconnected, roughly chronological essays. The first is the empire's struggle for unity in its diversity. The second is the problem of scale — an important element in climatology, where the range of interest runs from microclimates to global circulation.

The earlier Austrian Empire (1804–67) had suffered extensive territorial losses in a series of wars from the mid-eighteenth century to the Napoleonic era. Much of its energies were taken up with amalgamating the territory that remained. In 1867, a Habsburg compromise with the Hungarians created a larger, multi-ethnic collection of mainly Central European kingdoms under one imperial crown.

The new Austro-Hungarian Empire boasted more than ten major language groups, terrain ranging from alps to steppes, and transport and communication that remained rudimentary well into the nineteenth century. As Coen notes, successive governments under the long reign of Emperor Franz Josef I directed scholars to demonstrate that this all belonged under one flag. That was a potent issue in an age of national consolidation around linguistic units. The loose confederations of pre-unification Germany and Italy, for instance, could band together around a single national language and assertion of a common ethnicity. But this was not possible for the cultural mélange that was Austro-Hungary.

Universities, institutes, museums, herbaria, observational networks, publishing houses and government bureaus settled on climatology, meteorology and the metaphor of atmospheric circulation as the scientific proof of the 'naturalness' of the empire. This group developed a science designed to show the dynamic interdependence of regions with wildly diverse topography, hydrography and vegetation. Just as



ADALBERT STIFTER-GESELLSCHAFT, WIEN/WIEN MUS.

the wind from Austria brought rain to the Hungarian plain, and alpine snows fed the lands of the Danube, so each region was shown to provide some climatic essential that an adjoining one lacked.

These ideas infused society. They were picked up by economists including Emanuel Herrmann, who took climatology as a model for spatial analysis of the imperial economy. That concept, in turn, was developed by liberals such as the social democrat Karl Renner, who argued that diversity actually created unity, and that trade grew through exchange of excess goods between regions. Implicitly and explicitly, climatic interdependencies served as the foundation for political and economic oneness in a jostling, polyglot empire.

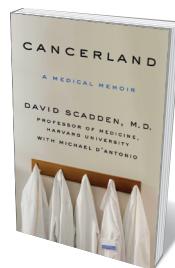
There was a closely linked emphasis on climate at different scales, micro to macro. Local topography and hydrography produce climates in areas ranging from a few kilometres to vast continental spans, and much of the work of climatology involves integrating these varied zones. But the focus on scale was driven by that overarching social, economic and political need for unity.

First, researchers studied conditions on very small scales, to satisfy local economic needs and political aspirations. These were then integrated into a larger imperial framework, especially in the data and maps of the huge, multi-volume series *Climatography of Austria*, published between 1904 and 1919 — a vastly more detailed descendant of German polymath Alexander von Humboldt's 1845 treatise *Cosmos*. Here, politics, rather than interfering, provided a rationale for doing a certain kind of science: descriptive, dynamic and focused on interdependence. That research was the seedbed of modern climatology. After the empire collapsed in 1918, catastrophically weakened by the First World War, the discipline was carried forward in Germany, Russia, Austria, Britain and North America.

There is a great deal more to this complex and reflective study. Coen examines how the empire promoted the popularization of science by leading experts while supporting research on the grand scale — an approach that stressed the patriotic, economic, cultural, even recreational utility of science. But the fact that climatology was born of a context of politics and policy, and was never far from them during its development, merits exactly this sort of examination as we wrestle with the ramifications of climate science today. ■

Mott Greene is affiliate professor of Earth and Space Sciences at the University of Washington in Seattle. He is the author of *Alfred Wegener: Science, Exploration and the Theory of Continental Drift*.
e-mail: mgreen@uw.edu

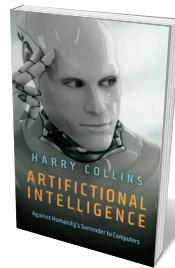
Books in brief



Cancerland: A Medical Memoir

David Scadden with Michael D'Antonio THOMAS DUNNE (2018)

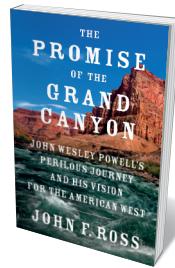
Cancer is a foreign country with different norms, asserts David Scadden in this gripping medical memoir. A leading immunologist and oncologist, Scadden (with writer Michael D'Antonio) examines the disease's recent history through interconnected lenses. Patients' often harrowing experiences twine through the narrative on research and treatments, from chemotherapy, bone-marrow transplants and lumpectomies to CRISPR and immunotherapies. Scadden's own eventful life in the lab (not least, his co-founding of the Harvard Stem Cell Institute in Cambridge, Massachusetts) is a highlight.



Artificial Intelligence

Harry Collins POLITY (2018)

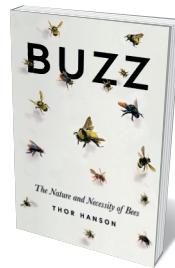
Sociologist of science Harry Collins has long focused on gravitational-wave physics (see *Nature* 542, 28–29; 2017). Here, he shifts gears dramatically to examine pervasive existential fears over artificial intelligence and its perceived threat in the 'deep learning' era. Collins probes this idea trenchantly and in considerable detail. Pointing to computers' inability to factor in social context, master natural language use well enough to pass a severe Turing test, or wield embodied cognition, he argues that the real danger we face is not a takeover by superior computers, but slavery to stupid ones.



The Promise of the Grand Canyon

John F. Ross VIKING (2018)

In 1869, US geologist John Wesley Powell was the first to explore the Grand Canyon and its environs scientifically in an intrepid river descent. Historian John Ross shows how, beyond the derring-do, Powell championed sustainable resource use and was a key architect of federal science. As head of the US Geological Survey, he created an important ecological map charting water scarcity in the West that aimed (but failed) to temper congressional dreams of manifest destiny. He even foresaw the 1930s Dust Bowl crisis. A bold study of an eco-visionary at a watershed moment in US history.



Buzz: The Nature and Necessity of Bees

Thor Hanson ICON (2018)

For this natural history of the bee, biologist Thor Hanson wings far beyond the hive to explore bee species from "bumbles" to wool carders. Here are the proto-bee of 125 million years ago, evolved from a Cretaceous wasp ancestor; a Chilean desert bee of the genus *Geodiscelis*, with a grotesquely elongated tongue; and a bumblebee colony's haphazard array of tiny wax pots. Here, too, are the data on dwindling populations. Apology, Hanson reminds us, is not just about the scientific buzz: bee behaviour has shed light on human issues from addiction to collective decision-making.



A Honeybee Heart Has Five Openings

Helen Jukes SCRIBNER (2018)

Joining the bright tide of cultural responses to all things apian (see *Nature* 521, 29–30; 2015) is this subtly wrought personal journey into the art and science of beekeeping. Helen Jukes evokes both the practical minutiae of the work, and the findings of researchers who have illuminated bee ethology over the centuries, from François Huber to Eva Crane. Laced through are quietly lyrical musings over 'hive life' that see Jukes perceiving her colony variously as a "brain with a million synapses", an inner citadel built by master architects or the fount of an only partly decoded language. *Barbara Kiser*

Poisonous politics in the Rust Belt

Mark Peplow extols two books on the water crisis in Flint, Michigan.

LeeAnne Walters and her family endured months of ill health before they discovered the source. In mid-2014, they developed skin rashes, lost clumps of hair and suffered mysterious aches. One of Walters's three-year-old twins stopped growing. By January 2015, the water supply in her home in Flint, Michigan, was brown. When she showed bottles of it to officials, they refused to believe it had come from her kitchen tap.

Flint's water was badly contaminated with lead, exposing tens of thousands of people to the potent neurotoxin. Amid denial and deception by authorities, a group of scientists, medics and engineers uncovered the scandal. But ultimately, the people of Flint turned the tide, thanks in part to remarkable citizen science that produced key water-quality data. Two books recount how the crisis unfolded.

The Poisoned City, by journalist Anna Clark, is gripping and packed with meticulously sourced reportage. *What the Eyes Don't See*, by Flint paediatrician Mona Hanna-Attisha, offers a powerful personal account of her role in the fight for justice.

Since the 1960s, Flint's water had come from Lake Huron by way of the Detroit Water and Sewerage Department (DWSD). But water rates were among the country's highest, so the cash-strapped city decided to set up its own provider. At first, it would take water from the Flint River and upgrade an old treatment plant. In April 2014, mayor Dayne Walling proudly switched on the supply.

Within weeks, residents' tap water began to taste metallic and smell rotten. As Clark and Hanna-Attisha reveal, the source was lead pipes that connected thousands of houses to the mains. Orthophosphate salts should have been added to the water to coat the pipes. But Flint's supply did not include this, breaching federal law. Chloride levels in the river water accelerated lead corrosion, which worsened when the treatment plant added ferric chloride to remove contaminants. The water also became tainted with bacteria, causing an outbreak of Legionnaire's disease.

For 18 months, officials insisted that the water met federal standards, and allegedly hid evidence. Some managers allegedly manipulated data to bring average lead levels below the regulatory limit of 15 parts per billion.

Clark's rich account intersperses policy and environmental science with vivid portraits of Flint and its citizens, ramping up the tension as the horror unfolds. She notes a turning point when Walters contacted environmental engineer Marc Edwards. Testing Walters's water in April 2015, he found lead levels



LeeAnne Walters holds a sample of her tap water.

hundreds of times those deemed acceptable: they averaged 2,000 p.p.b., with the highest more than 13,000 p.p.b.. So he mobilized an army of locals to collect samples.

His students distributed some 300 sampling kits, made an instructional video and set up a blog to report developments — a model of efficiency and transparency in marked contrast to how city, state and federal authorities acted. Yet when the team unveiled its findings in September 2015, officials dismissed the scientists, citizens and activists as rabble-rousers.

What finally forced the city to switch back to DWSD water a month later was proof that the supply was harming children. That came from Hanna-Attisha. She battled to access health records to show how levels of lead in children's blood had changed since the switch. She found that the proportion of under-fives with high blood lead levels had gone from 2.1% to 4%, topping 6% in the poorest areas.

Her book captures the urgency of dealing with a public-health emergency while maintaining rigour. She is honest about her fears of going public, anticipating that she would be vilified. Officials smeared her, distorted her findings and dismissed her evidence. A claim that she had "spliced and diced" the data hurt the most. "It felt like a public stoning," she writes, conveying the terror of a whistleblower confronting a powerful bureaucracy.

It would be all too easy to blame Flint's crisis on the incompetence of a few officials, but both authors pinpoint deeper factors. Flint is a classic Rust Belt city: its population has plummeted as industries shut plants. With fewer taxpayers, individual water bills soared, prompting the disastrous switch. Flint's leaky water-distribution network was

The Poisoned City: Flint's Water and the American Urban Tragedy

ANNA CLARK
Metropolitan (2018)

What the Eyes Don't See: A Story of Crisis, Resistance, and Hope in an American City

MONA HANNA-ATTISHA
OneWorld (2018)

also designed for a much larger population, making it even more expensive to maintain.

In 2011, a financial crisis prompted Michigan to appoint an 'emergency manager' to run the city, trumping the authority of mayor and city council. This led to a perfect storm of unaccountable decision-making, while budget cuts left environmental agencies poorly equipped to respond. Worse, the contamination particularly affected Flint's black residents, who tended to live in areas with the most poorly maintained water networks.

Clark and Hanna-Attisha identify these systemic failures — the impact of austerity, a breakdown in democracy and institutional racism — as the roots of Flint's water crisis. Their broader message is that these factors are not unique. Lead pipes are ubiquitous in US cities, and the Environmental Protection Agency has estimated that it would cost up to US\$80 billion to replace them. And chronic underfunding for public services continues to hit poor and minority communities hardest.

The story of Flint's crisis is still unfolding. Legal rulings have ordered the city to replace 18,000 pipes by 2020. Environmental and health officials are still on trial, facing charges including misconduct and tampering with evidence. It will be years before the health impact on Flint's children is fully understood.

These books, particularly Clark's, are must-reads — not only for those interested in environmental science and policymaking, but for anyone who believes that access to clean drinking water is a basic human right. ■

Mark Peplow is a science journalist based in Cambridge, UK.
e-mail: peplowscience@gmail.com

CORRECTIONS

The article 'Maria Mitchell at 200' (*Nature* **558**, 370–371; 2018) incorrectly said that Mitchell had a reflecting telescope at Vassar College; it was a refracting telescope.

In the review 'Israel's wild treasury' (*Nature* **558**, 516–517; 2018), the picture credit should have been "Itai Benit".

Correspondence

India must now work on renewables

In a landmark announcement, India's government declared in April that all of the 597,464 villages recorded in the national census now have access to electricity. Its next step should be to ensure that this energy comes from sustainable sources (see H. Nagendra *Nature* 557, 485–488; 2018).

The achievement is a significant advance in addressing global energy poverty: about 20% of the world's 1.1 billion people without access to electricity, and 30% of the 2.8 billion with no access to clean cooking energy, were in India (see go.nature.com/2iuftad). However, there is still some way to go. Up to 90% of villagers might still lack electricity, because a village is classified as electrified when 10% or more of its households are connected.

Another concern is the environment. India met 86% of its energy requirements from fossil sources in 2017–18 (64.8% from coal). More than one-quarter of the world's 6.5 million deaths due to air pollution in 2015 occurred in India.

More government investment in renewables is needed to ensure the long-term sustainability of India's electrical supply.

Shekhar Chandra, Lawrence E. Susskind *Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.*
shekharc@mit.edu

EPA transparency: open to scrutiny

I share Naomi Oreskes' concern that the US Environmental Protection Agency's (EPA's) new transparency rule is a disingenuous effort to discredit scientific findings and that it could prevent solid evidence from shaping regulations (*Nature* 557, 469; 2018). Other measures could boost transparency.

The EPA should improve the navigability of its website so that its decisions can be tracked and

scrutinized. It should also restore its library system and create public reading rooms in all of its 36 offices. And its administrator should reinvigorate existing transparency policies, such as the 2008 Quality Policy and the 2012 Scientific Integrity Policy.

Government and academic scientists have collaborated to develop a weight-of-evidence process to evaluate the available models and data (see <https://cfpub.epa.gov/si>). The resulting 'criteria documents' are comprehensively referenced and include details of the assessment and review procedures, as well as the assumptions, reference values and analytical parameters used. The process meets the requirements of the Clean Air Act, Clean Water Act and other federal statutes that define the EPA's mission. They are consistent with the Administrative Procedure Act and have been validated through state and federal court cases. Given their proven long-term track record, I see little value in extra administration protocols to address the transparency of decision-making.

Charles Herrick *Washington DC, USA.*
Ch133@nyu.edu

EPA transparency: justification for rule

As president of the US National Association of Scholars, I take issue with Naomi Oreskes' concerns over the transparency rule proposed by US Environmental Protection Agency (EPA) administrator Scott Pruitt (*Nature* 557, 469; 2018). In the association's view, the rule is a justified response to the irreproducibility crisis and reinforces the US government's long-standing commitment to base policy on the best available science.

In a public comment (see go.nature.com/2kugc81), the association recommends that the EPA should draft reproducible guidance to govern all of the

administrative processes involved in regulatory science. This document would define "best available science" as research that uses only pre-registered protocols and that provides data — along with associated protocols, computer codes, recorded factual materials and statistical analyses — that are archived and publicly available for continuing independent verification. Our proposed document should rescind the EPA's 'weight-of-evidence' standard for justifying regulatory policy and replace it with a "best available reproducible science" standard that also complies with that definition.

Peter Wood *National Association of Scholars, New York, USA.*
pwood@nas.org

University of Ghana, Legon-Accra, Ghana.

yaniweh@ug.edu.gh

**On behalf of 4 correspondents; see go.nature.com/2msdjjf*

Ancient molecule's 200th anniversary

It is 200 years since Louis Jacques Thenard discovered hydrogen peroxide by reacting barium peroxide with strong acids (L. Thenard *Ann. Chim. Phys.* 9, 314–317; 1818). Today, about 5 million tonnes of H_2O_2 is produced every year worldwide. Industry uses it as rocket fuel and a 'green' oxidant — for example, for treating wastewater and bleaching pulp and paper.

The molecule occurred in the oceans and in the atmosphere during prebiotic times, 4 billion years ago. At the time, there was no ozone layer and high-intensity ultraviolet irradiation generated the molecule through water radiolysis (J. Haqq-Misra *et al.* *Astrobiology* 11, 293–302; 2011). Early life forms soon developed specialized enzymes to break the molecule down into water and oxygen.

In the past 50 years, H_2O_2 attracted attention in molecular biology, after it was identified as a component of normal cell metabolism. High concentrations contribute to the inflammatory response and low concentrations have a signalling function (see, for example, H. Sies *et al.* *Annu. Rev. Biochem.* 86, 715–748; 2017).

This remarkable molecule fulfils the requirements for a biological messenger because it is relatively unreactive (W. H. Koppenol *et al.* *Free Radical Biol. Med.* 49, 317–322; 2010). Its enzymatic production and degradation, along with its ability to oxidize highly reactive protein thiol groups, equip it admirably for molecular signalling.

Willem H Koppenol *ETH Zurich, Switzerland*
Helmut Sies *Heinrich-Heine-University Düsseldorf*
koppenol@inorg.chem.ethz.ch

Living shapes engineered

Synthetic genetic circuits can induce cells to form simple 3D structures reminiscent of those generated during early embryonic development. This advance will help engineers build tissues that have desirable structures.

JESSE TORDOFF & RON WEISS

The structures of living organisms have properties that any engineer might hope to recreate. They can self-heal, grow and adapt, and they can have an astonishing range of material properties, from the strength of bone to the lightweight flexibility of an insect wing. To make these structures, a fertilized egg follows a developmental program — a set of instructions for cell behaviour encoded in its DNA. If we could understand and control the development of biological shapes, then we could harness the properties of living structures to build better organs *in vitro* and to generate designer materials that could mimic some of the abilities of living organisms. Writing in *Science*, Toda *et al.*¹ present a method for creating synthetic, designable developmental instructions, paving the way for researchers to engineer customizable biological shapes.

It has been proposed that all that is needed to make the diverse structures of the animal kingdom is a small set of fundamental tools — about ten shape-changing operations, including cell death, adhesion and movement². To decide which of these actions to use, cells can communicate with each other to establish their relative positions.

Toda *et al.* used an engineered cell-communication system called synNotch³ to mirror this biological set-up. SynNotch is adapted from Delta-Notch signalling — a signalling pathway found in nature, in which cells that have membrane-spanning Notch receptors sense Delta proteins on the surface of neighbouring cells. An intracellular effector domain is cleaved from Notch following ligand binding, and moves to the nucleus to regulate gene expression. In synNotch, the natural core of the Notch protein is used, but the ligand that is sensed and the effector domain that responds are customizable. In this way, it is possible to create multiple channels of modifiable cell-cell communication. With the appropriate choice of ligand and effector, the system can act independently of native Delta-Notch signalling to drive cell behaviour in customizable ways.

The authors engineered the cells so that the synNotch sensors regulated the expression of genes that encode cadherin proteins, which have long been known for their ability to create spatial organization in tissues. Cadherin

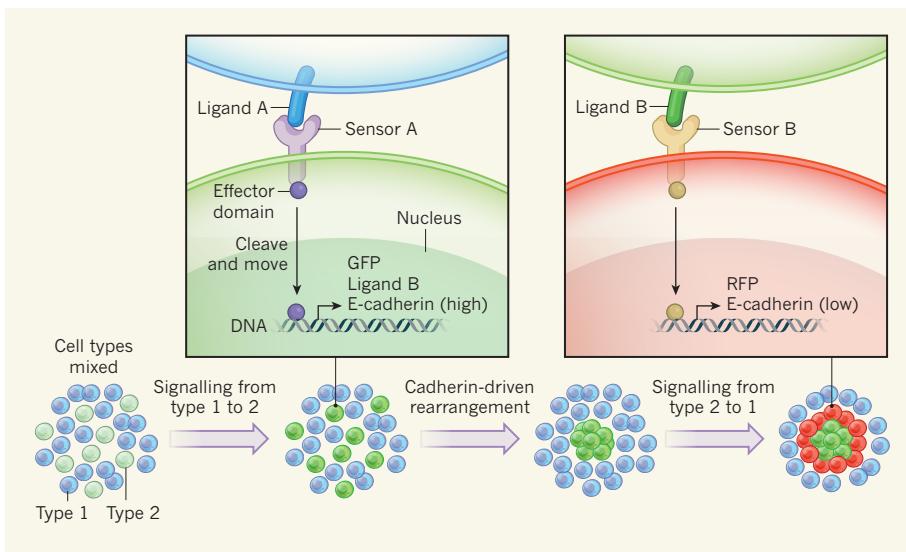


Figure 1 | Synthetic genetic circuits generate self-assembling structures. Toda *et al.*¹ have built circuits that, when expressed in different combinations in a population of cells, caused the cells to self-organize into various structures. In this example, the authors used two circuits. Initially, one group of cells (labelled type 1) expressed a blue fluorescent protein, and the other (type 2) did not fluoresce. When the populations were mixed, a ligand (designated A) produced by type 1 cells activated a receptor (sensor A) on type 2 cells, leading to cleavage of an intracellular effector domain from the sensor. This domain moved to the nucleus to trigger the expression of genes encoding ligand B, the adhesion protein E-cadherin and the protein GFP, which made the cells fluoresce green. E-cadherin caused type 2 cells to adhere to one another, rearranging the cell population. Ligand B then signalled to nearby type 1 cells, activating a different sensor (B). This led to the expression of a protein (RFP) that caused red fluorescence and low levels of E-cadherins. Because low E-cadherin production made the cells somewhat adhesive, they formed a second ring of cells. In this way, the circuits produced cycles of cell-cell communication and self-organization.

proteins mediate cell-cell adhesion, and so are essential for holding cells together and creating tissue boundaries during development⁴. Much like oil and water, cell populations that have different patterns or levels of cadherins can sort themselves into separate groups after being mixed together, and can self-assemble into a range of structures *in vitro*^{5,6}.

To create a synthetic program to guide shape formation, Toda *et al.* built several genetic circuits composed of different synNotch sensors that, when activated by a neighbouring cell, drive the expression not only of different levels or types of cadherin, but also of different ligands to bind to other sensors. In addition, each sensor drives the expression of a gene that encodes a fluorescent protein (green, red or blue), so changes in cell organization can be easily visualized. The authors mixed together cell populations harbouring these different circuits and allowed them to communicate

and move freely. They found that engineered communication between the cells led to cadherin-driven cell rearrangement, which in turn led to different cell-cell interactions, producing cycles of communication and shape change (Fig. 1).

Toda and colleagues observed remarkably complex cell behaviours. Cells self-organized to generate 3D structures, including a bullseye pattern and a sphere surrounded by multiple smaller nodes of different colours. The researchers could design instructions to produce specific structures, such as asymmetric forms — a key part of embryonic development. Furthermore, they showed that a structure of nested spheres could regenerate after being cut in two, as is often the case for self-organized tissues in living organisms.

The researchers next built a circuit to generate differential gene expression in a population of cells that was initially identical — a

process that mimics cell differentiation. To do this, they designed synNotch circuits to emulate one feature of the native Delta–Notch system known as lateral inhibition, in which Notch, when activated by Delta from a neighbouring cell, inhibits the expression of Delta in the receiving cell. This signalling produces a chequerboard pattern of two distinct cell populations, one expressing Notch, the other Delta, from an initially uniform population.

In the authors' lateral-inhibition circuit, one of these cell populations produced a green fluorescent protein, the other red. In addition, the two effector domains also promoted the production of different levels of the protein E-cadherin. In this way, the group was able to generate a structure that had rings of colour starting from a single uniform cell population.

With this work, Toda *et al.* have shown how we can design developmental programs to make new living shapes. Of course, there are limits to this approach. The authors' biggest structures are only a few hundred micrometres across, and adhesion-driven self-organization

alone is unlikely to generate structures of the size or complexity of organs. But advances in other types of synthetic-biology shape control could help to fill in some of the gaps. For instance, cells have been generated that can be artificially polarized such that asymmetric cell–cell contacts can be made⁷, and synthetic circuits have been designed to modify the behaviour of bacteria so that, across a whole population, arrangements are formed that resemble Turing patterns⁸. These patterns — such as stripes, spirals or the spots on a giraffe — arise during development as a result of biological signalling programs.

In the future, the toolkit established by Toda *et al.* could be expanded to generate short- and long-distance cell–cell communication alongside a synthetic system that controls all of the shape-changing operations involved in making biological structures. This could eventually give engineers total control when designing shapes that have some of the properties of living multicellular organisms. Such a development would be a huge advance. Not only could we map the rules of developmental biology by establishing

the limits and constraints of shape-changing biological operations, but we could also grow replacement organs and make adaptive living materials — for example, buildings that could construct and heal themselves. ■

Jesse Tordoff is in the Computational and Systems Biology Program, and **Ron Weiss** is in the Departments of Biological Engineering and of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
e-mails: tordoff@mit.edu; rweiss@mit.edu

1. Toda, S., Blauch, L. R., Tang, S. K. Y., Morsut, L. & Lim, W. A. *Science* **27**, eaat0271 (2018).
2. Davies, J. A. *J. Anat.* **212**, 707–719 (2008).
3. Morsut, L. *et al.* *Cell* **164**, 780–791 (2016).
4. Halbleib, J. M. & Nelson, W. J. *Genes Dev.* **20**, 3199–3214 (2006).
5. Nose, A., Nagafuchi, A. & Takeichi, M. *Cell* **54**, 993–1001 (1988).
6. Cachat, E. *et al.* *Sci. Rep.* **6**, 20664 (2016).
7. Loza, O. *et al.* *eLife* **6**, e24820 (2017).
8. Karig, D. *et al.* *Proc. Natl. Acad. Sci. USA* <https://doi.org/10.1073/pnas.1720770115> (2018).

This article was published online on 2 July 2018.

PALAECLIMATE

Hosing the North Pacific Ocean

Climate anomalies punctuated the last ice age, characterized by the discharge of icebergs that released fresh water into the North Atlantic Ocean. It now emerges that fresh water also sometimes flooded the North Pacific. **SEE LETTER P.241**

KAUSTUBH THIRUMALAI

Abrupt cold snaps known as Heinrich events occurred during past ice ages¹. These millennial-scale periods of colder climate were associated with massive influxes of fresh water to the North Atlantic Ocean. The influxes were caused by discharge of icebergs from the Laurentide Ice Sheet — an immense sheet of ice that covered most of central and eastern North America during glacial epochs (Fig. 1). Studies of such events are of great interest because they could help to indicate whether rapid reorganizations of ocean circulation might occur in the future, and how they might affect climate². On page 241, Maier *et al.*³ investigate how Heinrich events affected the North Pacific region during the last glaciation (roughly 115,000 to 12,000 years ago), and the influence of the Cordilleran Ice Sheet, North America's western counterpart of the Laurentide Ice Sheet. They report links between changes in ocean circulation in the North Atlantic and melting of the Cordilleran Ice Sheet.

There is abundant evidence that fleets of

icebergs episodically surged into the North Atlantic during the last glaciation. Heinrich events were initially identified from the coarse, ice-rafted detritus that forms layers in marine sediments¹. Numerous palaeoclimate records have since been obtained showing that ocean cooling and freshening (freshwater influx) occurred across the North Atlantic during Heinrich events⁴. The subsequent alteration of the Atlantic Ocean's circulation weakened heat transport between the hemispheres, and is hypothesized to have induced global temperature and precipitation anomalies through both atmospheric and oceanic pathways⁵.

It has been more challenging to find evidence that meltwater from the break-up of the Cordilleran Ice Sheet freshened the North Pacific Ocean. Near-coastal sediments in the northeast Pacific reveal that large abundances of freshwater biota were transported to the region by glacial-era meltwater⁶, and glacial debris has been uncovered in the region that can be associated with some Heinrich events⁷. By contrast, studies⁸ of planktic foraminifera (microscopic plankton that have shells made from calcium carbonate) preserved in



Figure 1 | Ancient ice sheets. During the last ice age, North America was covered by a complex of ice sheets, including the Laurentide Ice Sheet over the centre and east, and the Cordilleran Ice Sheet across the west; this map shows the maximum extent of the ice. Maier *et al.*³ analysed oxygen isotopes in the remains of organisms called diatoms trapped in sediments taken from the North Pacific Ocean (the star indicates the location of the sediment core studied). The changing ratio of isotopes in different layers of sediment reflects changes in the salinity of the sea water in which the diatoms lived. The isotopic measurements reveal that fresh water inundated the North Pacific during certain Heinrich events — millennial-scale periods during which the climate was anomalously cold. The authors conclude that the fresh water came from melting of the Cordilleran Ice Sheet.

sediments from the North Pacific indicate that no changes in salinity occurred during North Atlantic Heinrich events, muddying the picture of how the Cordilleran Ice Sheet affected ocean dynamics and climate during these events.

Maier *et al.* now report a study of marine diatoms — single-celled plankton that have

shells made from silica — preserved in open-ocean sediments from the northeastern North Pacific (Fig. 1). The authors measured the ratios of stable oxygen isotopes in the diatoms. These ratios reflect past changes in the temperature and isotopic composition of sea water, which, in turn, vary with changes in global sea level and local salinity. The same principle was used in earlier studies⁸ of planktic foraminifera, but diatoms can thrive in colder and less saline environments⁹ than can many planktic foraminiferal species. Maier and colleagues' measurements reveal that large and abrupt intrusions of low-salinity waters occurred at their study site, coinciding with the timing of some Heinrich events. The authors interpreted these intrusions as evidence of meltwater originating from the Cordilleran Ice Sheet.

The researchers went on to carry out a series of computational climate-modelling experiments, of a type known as hosing experiments. Such simulations are used to study anomalies in global climate and ocean circulation that arise in response to abrupt climate change, and involve artificially introducing fresh water into the oceans at high latitudes, typically routed to the North Atlantic⁵. Maier *et al.* extended the hosing approach in two ways. First, they used a climate model that represents isotopic tracers, which thus enabled a more direct comparison of the simulations with their measurements. And second, they performed two sets of simulations, one in which only the North Atlantic was hosed, and the other in which both the North Atlantic and North Pacific were hosed.

The authors found that the simulation in which freshwater input was confined to the North Atlantic did not indicate that low-salinity waters entered the North Pacific Ocean, contradicting the findings from their diatom measurements. Despite this difference, the simulation did reproduce the ocean–atmosphere dynamics thought to have occurred across the Pacific in response to perturbations in the North Atlantic⁵. It also revealed poleward routing of warm, subtropical ocean waters due to shifts in tropical rainfall. Maier *et al.* therefore propose that the rerouted warm waters might have been responsible for the melting of parts of the Cordilleran Ice Sheet, adding fresh water to the North Pacific — a scenario that they could model by hosing the North Pacific as well as the North Atlantic.

Indeed, when the authors simulated this scenario, it provided a better match to the diatom observations. Moreover, the simulation suggests that salinity changed only negligibly at depth in the North Pacific. This might explain why no change in salinity was recorded in the isotopic study of foraminifera — it is thought that these organisms do not dwell at the topmost part of the ocean in this region. However, the diatom data indicate that freshwater influxes to the North Pacific did not occur during every Heinrich event. This could be because all Heinrich events are not created equally^{1,4}. Sure enough, when Maier and

colleagues performed additional simulations of North Atlantic perturbations involving exceptionally cool background temperatures, they found that the conditions produced were not conducive to melting of the Cordilleran Ice Sheet.

The new study is a major advance in our understanding of freshwater events in the North Pacific, but questions remain owing to the limitations of the time resolution of the sediments in the core that was analysed, and because the low abundance of diatoms in some sedimentary layers prevented the authors from carrying out their analysis for the corresponding periods of geological time. For example, the lack of evidence of freshwater pulses in proxies of the North Pacific surface ocean during some Heinrich events is puzzling, given the presence of glacial detritus. Furthermore, we still do not know how stable the Cordilleran Ice Sheet would be in response to shifts in Pacific climate that are unrelated to Heinrich events.

Importantly, further research is required to determine whether Cordilleran-meltwater

events influenced circulation in the Pacific, or even in the Atlantic. More broadly, a more-refined understanding of Cordilleran-meltwater pulses and the associated effects on regional temperature and precipitation will benefit our theories of abrupt climate change. ■

Kaustubh Thirumalai is in the Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, Rhode Island 02912, USA.
e-mail: kaustubh_thirumalai@brown.edu

1. Heinrich, H. *Quat. Res.* **29**, 142–152 (1988).
2. Liu, W., Xie, S.-P., Liu, Z. & Zhu, J. *Sci. Adv.* **3**, e1601666 (2017).
3. Maier, E. *et al.* *Nature* **559**, 241–245 (2018).
4. Hemming, S. R. *Rev. Geophys.* **42**, RG000128 (2004).
5. Okumura, Y. M., Deser, C., Hu, A., Timmermann, A. & Xie, S.-P. *J. Clim.* **22**, 1424–1445 (2009).
6. Lopes, C. & Mix, A. C. *Geology* **37**, 79–82 (2009).
7. Hendy, I. L. & Cosma, T. *Paleoceanography* **23**, PA2101 (2008).
8. Maier, E. *et al.* *Paleoceanography* **30**, 949–968 (2015).
9. Harrison, P. J. *et al.* *Prog. Oceanogr.* **43**, 205–234 (1999).

DNA DAMAGE

Breaking the replication speed limit

Inhibitors of PARP proteins are used in cancer treatment. It emerges that PARP inhibitors exert their effect by accelerating DNA replication to a speed at which DNA damage occurs. SEE LETTER P.279

ANNABEL QUINET & ALESSANDRO VINDIGNI

The two strands of DNA's double helix unwind to be copied, with a structure called a replication fork forming at the point of separation. The speed at which the replication fork progresses along DNA — and so the speed of replication — must be controlled to guarantee faithful duplication of the genome. On page 279, Maya-Mendoza *et al.*¹ define a molecular network involved in the regulation of replication-fork speed. Changes to this network can cause that speed to increase above a safe threshold, causing DNA damage and genomic instability.

Replication forks that encounter damage in the genome sometimes temporarily stop, allowing DNA repair to occur before replication continues. Proteins of the poly(ADP) ribose polymerase (PARP) family, particularly PARP1, assist in the repair of breaks in single strands of DNA through a process called PARylation², in which the proteins synthesize chains of ADP-ribose molecules that attract repair proteins to the damaged DNA. PARP inhibitors — drugs that block the PARylation activity of PARP proteins — are

showing promise as therapeutics to treat various cancer types³. Previous models have proposed that, by preventing PARP activity, PARP inhibitors cause replication forks to stall for abnormally long periods, and eventually to collapse, when they encounter DNA damage⁴. This leads to accumulation of DNA damage owing to improper replication and death of the treated cells⁴.

Maya-Mendoza *et al.* challenge the idea that PARP inhibitors perturb the ability of replication forks to progress. The authors found that treating proliferating human cells with the PARP inhibitor olaparib *in vitro* led to aberrant acceleration of fork speed. They provide evidence that, if fork speed increases above a threshold speed of 40% faster than normal, there is insufficient time for the forks to recognize damaged DNA in need of repair. This leads to accumulation of DNA damage and reduced cell viability. Supporting this idea, the authors found that violation of the threshold speed led to the activation of proteins involved in a DNA-damage response, although the mechanism by which this occurs needs to be further investigated.

To uncover the pathway by which PARP

inhibition speeds up replication forks. Maya-Mendoza *et al.* investigated the protein p21, which can inhibit DNA replication⁵. Expression of the *p21* gene is controlled by PARP1 (ref. 6). Moreover, the protein p53, which is a central player in maintenance of genome integrity, activates expression of *p21* (ref. 7) and is itself activated by PARylation⁸. The authors found that loss of *p21* led to an increase in replication-fork speed similar to the acceleration caused by PARP inhibitors. Loss of *p21* in addition to PARP inhibition increased fork speed more than either manipulation in isolation. Combining these observations, the authors propose the existence of a fork-speed regulatory network that has two interacting arms — the p53–p21 axis and PARylation. Each arm acts to keep fork speed below the threshold, with inhibition of either p21 or PARylation throwing the network out of balance and so increasing fork speed (Fig. 1). Several steps of this regulatory pathway will require further investigation. For example, exactly how the arms interact to properly control replication-fork speed is a key question to address.

PARP inhibitors are used to treat tumours that have deficiencies in a pathway called homologous repair that repairs double-stranded DNA breaks. These defects make the tumour cells particularly susceptible to PARP inhibitors. To explain, the single-stranded DNA breaks that accumulate owing to PARP inhibitors are converted to double-stranded breaks when the damaged strand is replicated. In normal cells, the breaks can then be repaired through homologous repair, but when this pathway is defective, the inability to repair these defects leads to cell death. Most notably, breast, ovarian and prostate cancers caused by mutations in the genes *BRCA1* and *BRCA2* are susceptible to PARP inhibition^{9,10}.

Maya-Mendoza *et al.* found that PARP inhibition accelerates fork speed above the threshold in *BRCA1*-deficient cells. On the basis of these results, the authors suggest that the susceptibility of tumours harbouring *BRCA* mutations to PARP inhibitors might not be due to increased stalling and collapse of replication forks, as originally believed, but instead to aberrant acceleration that compromises the ability of forks to detect and repair DNA damage.

In summary, Maya-Mendoza *et al.* have provided a fresh view of why PARP inhibitors are toxic to cancer cells, and have outlined a previously unknown network that controls replication-fork speed. Their work has the potential to revolutionize current models of how cells cope with DNA damage — but it also raises several questions.

For example, do other PARP proteins help to control fork speed? The authors report that PARylation levels were not affected by PARP1 depletion. This observation implies that other members of the PARP family are involved in controlling replication-fork speed.

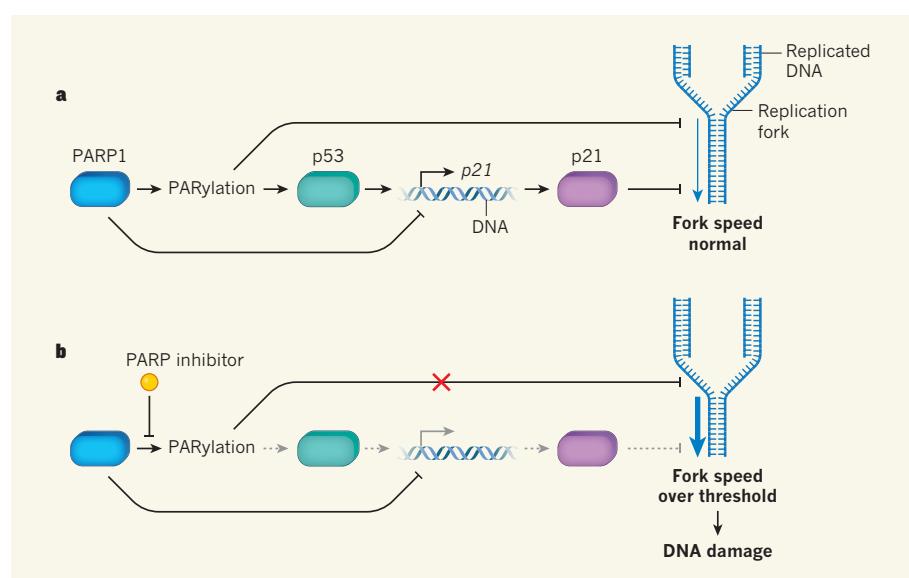


Figure 1 | A regulatory network controlling replication-fork speed. DNA undergoing replication unwinds at dynamic structures called replication forks, which move through the genome as replication progresses. Maya-Mendoza *et al.*¹ have unravelled a two-armed regulatory network that controls replication-fork speed. **a**, In one arm, the protein PARP1 limits fork speed through an enzymatic process called PARylation. In the other, the protein p53 activates the gene *p21*, which encodes a protein that also limits fork speed. The two arms are interconnected because PARylation regulates p53 activity and PARP1 inhibits *p21* expression. **b**, Drugs called PARP inhibitors, which prevent PARylation, prevent the network from properly inhibiting replication-fork speed, although the effect of the drugs on the p53–p21 arm remains to be fully elucidated (indicated by faded, dashed arrows). Fork speed rises to more than 40% faster than normal, and crossing this threshold leads to DNA damage and genomic instability.

By what mechanism do PARP activity and p21 control replication-fork speed? PARP activity is crucial in the control of replication-fork reversal — a mechanism by which replication forks reverse their course when they face DNA breaks^{11,12}, enabling the damage to be dealt with. Perhaps PARP activity and p21 affect fork speed by suppressing replication-fork reversal or other mechanisms used by replication forks to cope with DNA breaks.

There are other areas of interest for future

“Maya-Mendoza et al. have provided a fresh view of why PARP inhibitors are toxic to cancer cells.”

research. For example, the effect of increased fork speed on polymerase enzymes, which carry out DNA replication, should be examined to determine whether the enzymes exacerbate the situation by introducing more errors into the newly replicated genome as a consequence of increased fork speed. Whether the toxic effects of PARP inhibitors on cancer cells are mainly linked to the fact that the forks do not detect damage when the threshold speed is violated remains to be confirmed.

Notably, PARP inhibitors have also been effectively used in combination with chemotherapeutic agents, which induce DNA damage by impairing the ability of replication forks to progress. It would therefore be interesting to determine whether the same mechanism

underlies the effects of PARP inhibitors when used in combination with chemotherapy. Finally, Maya-Mendoza and colleagues' findings will no doubt prompt many research groups to explore whether surpassing the threshold fork speed provides a more general way to explain the molecular basis of cancer and tumour sensitivity to chemotherapy. ■

Annabel Quinet and Alessandro Vindigni are in the Edward A. Doisy Department of Biochemistry and Molecular Biology, Saint Louis University School of Medicine, St. Louis, Missouri 63104, USA.
e-mails: annabel.quinetdeandrade@health.slu.edu; alessandro.vindigni@health.slu.edu

1. Maya-Mendoza, A. *et al.* *Nature* **559**, 279–284 (2018).
2. Gibson, B. A. & Kraus, W. L. *Nature Rev. Mol. Cell Biol.* **13**, 411–424 (2012).
3. Lord, C. J. & Ashworth, A. *Science* **355**, 1152–1158 (2017).
4. Bryant, H. E. *et al.* *EMBO J.* **28**, 2601–2615 (2009).
5. Waga, S., Hannon, G. J., Beach, D. & Stillman, B. *Nature* **369**, 574–578 (1994).
6. Madison, D. L. & Lundblad, J. R. *Oncogene* **29**, 6027–6039 (2010).
7. el-Deiry, W. S. *et al.* *Cell* **75**, 817–825 (1993).
8. Lee, M. H., Na, H., Kim, E. J., Lee, H. W. & Lee, M. O. *Oncogene* **31**, 5099–5107 (2012).
9. Bryant, H. E. *et al.* *Nature* **434**, 913–917 (2005).
10. Farmer, H. *et al.* *Nature* **434**, 917–921 (2005).
11. Berti, M. & Vindigni, A. *Nature Struct. Mol. Biol.* **23**, 103–109 (2016).
12. Ray Chaudhuri, A. *et al.* *Nature Struct. Mol. Biol.* **19**, 417–423 (2012).

This article was published online on 27 June 2018.

Infant deaths from air pollution estimated

A carefully considered observational study estimates that up to 22% of infant deaths in sub-Saharan Africa could be prevented by improving air quality — a value much higher than previous estimates. [SEE LETTER P.254](#)

LANCE A. WALLER

Big data can help to address many pervasive problems in the field of public health. For instance, large-scale data analyses are helping researchers to understand global patterns of disease, the range of factors that contribute to global health and the policies that provide the greatest potential for improvement^{1,2}. On page 254, Heft-Neal *et al.*³ propose, implement and (importantly) scrutinize such an approach, exploring the impact of air quality on infant mortality in sub-Saharan Africa.

The authors' study joins a growing body of work that explores international patterns of health outcomes through creative analyses of big data — a set of approaches pioneered by many on local geographical scales, but brought to the global-health stage by a project called the Global Burden of Disease Study (GBDS). In these types of study, multiple sources of health, administrative and research data are pooled and subjected to mathematical modelling and complex statistical analysis. But this exciting branch of public-health research is still finding its place amid conventional epidemiological techniques that involve gathering data from direct observations in cases and controls, or in longitudinal studies.

GBDS data have previously been used to estimate links between local air quality and mortality on a global scale (for example, in the project's 2016 report⁴). But these analyses were dominated by data obtained from air-pollution monitoring stations, which are predominantly found in developed countries. In these areas, air pollution is typically lower than in sub-Saharan Africa.

By contrast, Heft-Neal *et al.* used satellite-based measurements of air pollution. They combined these measurements with data from 65 household health surveys, which they used to determine mortality for almost 1 million births in 30 countries across sub-Saharan Africa between 2001 and 2015. The authors also focused on infant mortality from all causes, whereas the GBDS emphasized mortality due to respiratory illness.

The results are surprising. Heft-Neal *et al.* estimate that 22% of infant deaths in sub-Saharan Africa — a total of 449,000 — could be avoided by decreasing average levels of

air pollution to the lowest levels observed in the region (a concentration of 2 micrograms per cubic metre). This level of comparative improvement is higher than the estimates reached by two previous analyses using the publicly available GBDS data^{5,6} (Fig. 1). The authors place their results in the context of the previous work, putting forward several reasons for the different values. These include differing assumptions about what level of improvement in air quality is attainable (improvement from a median of 25 to 2 $\mu\text{g m}^{-3}$ in the present paper, compared with improvement to 5.8 $\mu\text{g m}^{-3}$ in the earlier analyses) and different sets of mortality data.

Rather than being satisfied with the headline association alone, Heft-Neal and colleagues carefully review the uncertainty in

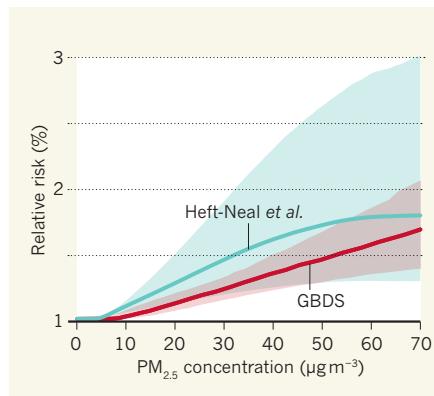


Figure 1 | Two estimates of the risk of infant death linked to air pollution. Air pollution can be measured as a concentration of breathable particulate matter ($\text{PM}_{2.5}$) in micrograms per cubic metre ($\mu\text{g m}^{-3}$). The Global Burden of Disease Study (GBDS)⁶ estimated the relationship between increasing $\text{PM}_{2.5}$ and relative risk of infant mortality due to respiratory infections globally. By contrast, Heft-Neal *et al.*³ used different data-analysis approaches to estimate the relative risk of all-cause infant mortality related to air pollution only in sub-Saharan Africa, where pollution rates are generally higher than in wealthier regions of the world. The general message is the same (a clear benefit from reducing levels of air pollution), but Heft-Neal *et al.* find a greater increase in mortality with increasing air pollution. The levels of uncertainty (shaded areas) provide essential context for understanding the results. (Adapted from Fig. 3 of ref. 3.)

their estimation. For instance, they detail how the results might be affected by analytical assumptions, such as a linear relationship between air pollution and mortality within the range of observed values, and potential biases associated with using satellite-based measurements as a proxy for air pollution at ground level. They also consider potential confounders such as socio-economic status — it has previously been predicted that wealthier households would be less affected by air pollution than poorer households, but the authors show that this is not the case in their analysis. Such self-reflection is refreshing and essential, and places the results in an appropriate context for consideration by researchers and policy experts.

Heft-Neal *et al.* outline their data sources in their supplementary information, but future work can go further by filling in the details necessary to replicate and reproduce results from big-data studies. For example, detailed, peer-reviewed descriptions of data curation should be published, and the final data set should itself be deposited in citable repositories such as datadryad.org. By sharing citable analysis details and data, the value of studies such as Heft-Neal and colleagues' could be even greater.

Is this the final word on associations between air quality and infant mortality? Certainly not, because any observational study runs the risk of confusing correlation with causation. But I would suggest that proof of causation should not be the only motivation for such studies. Rather, the goal of any scientific exploration should be to know more afterwards than we did before. Proving causation might help researchers to pinpoint the direct effects of particular policies on particular aspects of health. But carefully vetted broad-scale associations can point to ways in which small policy changes can yield large improvements (even if indirectly) in addressing challenging public-health goals. This is especially useful for aspects of public health, such as air-pollution analyses, in which tightly controlled experimental studies would be difficult and ethically challenging — it would not be possible, for instance, to randomize levels of air pollution to individuals, nor to easily assign specific exposures to specific locations.

Large-scale data-science studies can offer insight into factors that predict trends in health outcomes, but may have limited use for defining causation, particularly at continental scales. For example, consider Google Flu, which aimed to estimate the numbers of influenza cases in the United States by analysing search-term trends relating to flu symptoms. For many weeks, the system's data-science-based predictive approach provided more-accurate results than did conventional epidemiological tracking based on physician reports and laboratory confirmation. However, following an adjustment to the prediction algorithm in

early 2013, the system vastly overestimated flu cases for two weeks⁷. By relying wholly on associations rather than also incorporating epidemiological risk factors, the algorithm had few checks and balances against over- or underestimation, and offered few insights into the factors driving short-term patterns in flu incidence.

In summary, although big-data analyses cannot replace careful epidemiological studies, they can give broad insight into the potential benefits of public-health policies. In this case, Heft-Neal and colleagues' work highlights the benefits of aspiring to reduce air pollution to the lowest levels observed in their data set, and provides assessments of the effects of more-modest changes in pollution levels. This type of analysis certainly has a place in the modern public-health toolbox. As noted by

Kofi Annan²: "Without good data, we're flying blind. If you can't see it, you can't solve it." ■

Lance A. Waller is in the Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322, USA.

e-mail: lwaller@emory.edu

1. Dowell, S. F., Blazes, D. & Desmond-Hellmann, S. *Nature* **540**, 189–191 (2016).
2. Annan, K. *Nature* **555**, 7 (2018).
3. Heft-Neal, S., Burney, J., Bendavid, E. & Burke, M. *Nature* **559**, 254–258 (2018).
4. GBD 2016 Mortality Collaborators. *Lancet* **390**, 1084–1150 (2017).
5. Cohen, A. J. *et al.* *Lancet* **389**, 1907–1918 (2017).
6. Burnett, R. T. *et al.* *Environ. Health Perspect.* **122**, 397–403 (2014).
7. Lazer, D., Kennedy, R., King, G. & Vespignani, A. *Science* **343**, 1203–1205 (2014).

This article was published online on 27 June 2018.

CONDENSED-MATTER PHYSICS

The heat is on for Majorana fermions

Exotic particles called Majorana fermions have potential applications in quantum computing, but their existence has yet to be definitively confirmed. Two groups have now glimpsed these particles. **SEE ARTICLE P.205 & LETTER P.227**

KIRILL SHTENGEL

The building-blocks of matter — protons, neutrons and electrons — are examples of particles called fermions. Eight decades ago, the Italian physicist Ettore Majorana predicted the existence of fermions that are their own antiparticles¹. These particles, now known as Majorana fermions, would be of great fundamental interest, and could revolutionize quantum computing. Evidence for Majorana fermions among elementary particles remains elusive; however, in the past few years, there has been striking progress in this hunt in the realm of condensed-matter physics². On pages 205 and 227, respectively, Banerjee *et al.*³ and Kasahara *et al.*⁴ report signatures of Majorana fermions in heat-transport experiments in two very different condensed-matter settings.

Condensed-matter systems contain excitations that behave like ordinary particles, but that need not resemble the actual elementary particles that the systems are made of. For example, the phenomenon of superconductivity (more specifically, topological superconductivity) provides a setting in which an electron can effectively 'forget' its electric charge. As a result, the electron becomes indistinguishable from its antiparticle, which in this context is an electron vacancy called a hole. Whether topological superconductivity

is an intrinsic feature of solid-state materials remains an open question. However, the key aspects of the phenomenon can be mimicked in certain condensed-matter systems, providing the right conditions for the emergence of Majorana fermions. The two systems investigated in the current papers seem to be of just this kind.

Banerjee and colleagues looked for evidence of Majorana fermions on the edge of a condensed-matter system that exhibits the quantum Hall effect — whereby, at low temperature and in the presence of a strong magnetic field, the material's transverse electrical conductance becomes quantized (it can have only specific values). The authors focused on a particular state for which this conductance is 5/2 times the fundamental unit. The exact nature of this state has been a subject of debate, but all of the strong contenders can be thought of as superconducting states of composite fermions⁵.

By contrast, Kasahara and colleagues investigated a form of ruthenium chloride known as α -RuCl₃. This material is thought to be in a phase known as the Kitaev spin liquid — a peculiar state of matter that lacks long-range magnetic order all the way down to zero kelvin^{6,7}. Although α -RuCl₃ is an electrical insulator, the description of the magnetic properties of a Kitaev spin liquid is mathematically equivalent to that of a topological



50 Years Ago

Motorists in south and central England who cleaned their cars during the last weekend of June regretted their diligence when they rose on Monday, July 1, to find deposits of orange coloured dust over every exposed surface. The explanation was an early morning shower of rain, laden with dust swept up probably from somewhere in North Africa ... this was an unusual event, even for a country which prides itself on the peculiarities of its climate ... it turned out that the last time a dust fall like this happened was in 1903 ... On the same day as the widespread dust fall, Minehead ... Dulverton ... and Burnley ... were bombarded by hailstones the size of golf balls ... The Meteorological Office is keeping an open mind on whether this is a coincidence or whether there was a causal relationship between the dust and the hail.

From *Nature* 13 July 1968

100 Years Ago

During the last twenty years there has been an extraordinary increase in the ... output of books and papers on scientific subjects. In the olden time many a quiet student would be content to spend his life upon one piece of work ... in the hope that it might remain a permanent addition to human knowledge ... [A]nyone wishing to learn the present state of our knowledge ... might well despair of ever discovering all that has recently been written ... A complete catalogue of all scientific publications throughout the world would be, unfortunately, very bulky ... An alternative method is to draw up a list of journals ... and to confine the catalogue to papers published in these journals. When this plan is adopted it is hoped that authors ... will gradually acquire the habit of sending any original paper they wish to publish to one of these periodicals.

From *Nature* 11 July 1918

superconductor. Therefore, Majorana fermions should exist on the edge of α -RuCl₃.

The direct detection of Majorana fermions in condensed-matter systems was never going to be easy. Such particles must be electrically neutral and therefore cannot participate in electrical transport (although they can mediate such transport in superconductors^{8,9}). However, although Majorana fermions are unable to conduct current, they can conduct heat.

Electrons can conduct both electricity and heat. As a result, metals — which contain many free electrons — are typically good heat conductors. This idea is formalized by the Wiedemann–Franz law, which states that electrical conductivity is directly proportional to thermal conductivity divided by temperature. Although the identification of this relationship is often lauded as one of the early successes of solid-state theory, the proportionality constant is not universal for ordinary metals: scattering processes, which limit both electrical and thermal conductivity, affect these properties differently in different metals.

However, if the motion of particles in a material is ballistic (if there is effectively no scattering), both electrical and thermal conductivity are quantized and proportional to the number of propagating modes (conduction channels). Each electron mode contributes a unit of thermal conductance, and, crucially, each Majorana mode contributes only half a unit. Both Banerjee *et al.* and Kasahara *et al.* observed this fraction of thermal conductance on the edges of their condensed-matter systems.

The existence of Majorana edge modes in a condensed-matter system is a strong indicator that the topological order of the system is non-Abelian — which means, for example, that a collection of the system's excitations has a huge number of quantum states with the same energy. The non-Abelian nature of the quantum Hall state studied by Banerjee *et al.* has long been expected (albeit not confirmed beyond reasonable doubt). However, Kasahara and colleagues' findings provide the first experimental evidence of a non-Abelian spin liquid. Although more work is needed to confirm the exact nature of this state, the discovery of such an unconventional phase of matter is truly exciting.

Banerjee and colleagues used their measurements to try to discriminate between different candidate non-Abelian states. This task is harder than obtaining evidence for non-Abelian topological order. It relies on counting both fractional and integer contributions to the system's thermal conductance, which, in turn, requires certain assumptions to be made about the process by which different propagating modes reach thermal equilibrium¹⁰. The issue of equilibration is further complicated by the fact that the edge modes can reach equilibrium not only with each other, but also with lattice vibrations

called phonons, which provide an unwanted contribution to the thermal conductance.

Banerjee *et al.* went to great lengths to minimize this phonon contribution. They carried out their experiments at temperatures of about 20 mK and used a sophisticated design of a source and drains to avoid the coupling of edge modes to phonons. By comparison, Kasahara and colleagues' experiment was much less intricate and required temperatures of only about 5 K. These authors could not detect a signal of half-integer quantization at lower temperatures, which probably suggests that the system transitioned to a different phase. Their results also indicate that a substantial amount of heat was carried by phonons.

Under these circumstances, it should be surprising that the authors saw signs of quantized Hall heat transport — the heat conduction in the direction perpendicular to that of the thermal gradient — by Majorana fermions. However, two recent studies^{11,12} have argued that phonon coupling not only is not detrimental, but also can actually be necessary for

the observation of such an effect. More work, both theoretical and experimental, is required to fully understand the implications of these experiments. Nevertheless, it is undoubtedly exciting that the quest for Majorana fermions is heating up in this manner. ■

Kirill Shtengel is in the Department of Physics and Astronomy, University of California, Riverside, California 92521, USA. e-mail: kirill.shtengel@ucr.edu

1. Majorana, E. *Nuovo Cimento* **14**, 171–184 (1937).
2. Wilczek, F. *Nature Phys.* **5**, 614–618 (2009).
3. Banerjee, M. *et al.* *Nature* **559**, 205–210 (2018).
4. Kasahara, Y. *et al.* *Nature* **559**, 227–231 (2018).
5. Read, N. & Green, D. *Phys. Rev. B* **61**, 10267–10297 (2000).
6. Kitaev, A. *Ann. Phys.* **321**, 2–111 (2006).
7. Jackeli, G. & Khaliullin, G. *Phys. Rev. Lett.* **102**, 017205 (2009).
8. Mourik, V. *et al.* *Science* **336**, 1003–1007 (2012).
9. He, Q. L. *et al.* *Science* **357**, 294–299 (2017).
10. Simon S. H. *Phys. Rev. B* **97**, 121406 (2018).
11. Ye, M., Halász, G. B., Savary, L. & Balents, L. Preprint at <https://arxiv.org/abs/1805.10532> (2018).
12. Vinkler-Aviv, Y. & Rosch, A. Preprint at <https://arxiv.org/abs/1805.11587> (2018).

ECOLOGY

How rats wreak havoc on coral reefs

The introduction of non-native rats can devastate island ecosystems. It now emerges that these rats also harm a complex web of interactions linking seabirds with the algae and fishes of nearby coral reefs. SEE LETTER P250

NANCY KNOWLTON

Non-native rats that invade tropical islands can cause problems for the ecosystems they invade¹. These intruders can decimate the native populations on which they feed, such as plants and terrestrial invertebrates. Bird populations can plummet, too, when rats eat eggs and nestlings. The complex, indirect effects of rodent presence can spread deeply and widely through island food webs². However, little attention has been paid to the indirect impacts of such invasive species on adjacent coral-reef communities. On page 250, Graham *et al.*³ address this for sites on the Chagos Archipelago in the Indian Ocean (Fig. 1), comparing coral reefs surrounding six rat-infested islands with those adjacent to six islands that lacked rats. The authors find that reefs near rat-infested islands have fewer nutrients, fewer fishes and reduced numbers of fishes grazing on the algae that compete with corals.

One of the most marked effects of rats was a 760-fold decline in the number of nesting seabirds per hectare on rat-infested islands compared with rat-free islands. On the

latter islands, the larger populations of birds produced larger deposits of guano — nitrogen-rich bird excrement. This nitrogen is mostly obtained from food that the birds consume during long-distance foraging trips to parts of the ocean that, thanks to their higher levels of nutrients, are 100 to 100,000 times more productive than the waters in the immediate vicinity of an island. The nitrogen deposition rates on the rat-free islands were 251 times greater per hectare than were those on the rat-infested islands. Using a technique to identify different isotopic forms of nitrogen, the authors could distinguish this 'imported' seabird nitrogen from locally derived nitrogen. This enabled Graham and colleagues to track where the seabird-deposited nitrogen ended up.

Some nitrogen was absorbed by plants on the islands, and some entered the ocean through rain or breaking waves. For example, 100 metres from the shore of rat-free islands, both a type of sponge and a type of macroalgae had elevated levels of nitrogen derived from seabird foraging, compared with the levels recorded near rat-infested islands. At 230 metres from the shore of rat-free islands, the concentration of seabird-derived nitrogen

NICK GRAHAM in turf algae and in the muscle tissue of a species of alga-eating damselfish was higher than such measurements on rat-infested islands. By measuring growth rings in damselfish ear bones, Graham and colleagues showed that the damselfish in the waters around rat-free islands grew faster and at any given age were larger than those living beside rat-infested islands, presumably because their food was richer in nitrogen.

Looking at all types of reef fish, the authors found that the total biomass of the fish population was 48% higher around rat-free islands than around rat-infested islands. Moreover, of all the types of reef fish, the abundance of herbivorous (alga-eating) fishes was the most negatively affected by the presence of rats. Herbivorous fish are particularly important for coral reefs, because their grazing prevents the algae from overgrowing and killing corals. Around rat-free islands, parrotfishes, a group of herbivores, grazed the entire surface of the reef 9 times per year, whereas around rat-infested islands the equivalent figure was only 2.8 times per year. Because parrotfishes feed with powerful beaks, there was also more bioerosion and greater production of sand on coral reefs surrounding rat-free islands; however, the amount of living coral was not lower than that on rat-infested islands.

The dramatic effects that Graham and colleagues document provide a comprehensive picture of how reefs are interconnected with the surrounding land and seascapes. The movement of organisms around such habitats generates genetic connections between regions that have long been appreciated by biogeographers and geneticists. In marine conservation, the generation of networks of marine-protected areas that take into account this genetic connectivity between coral reefs is increasingly a part of the planning process. John Donne's poem 'No Man is an Island' explores the nature of human connections, and, as others have noted⁴ in a similar vein, "no reef is an island", either.

However, the nutritional (trophic) connections linking coral reefs with other marine and terrestrial ecosystems haven't been studied as extensively as have the genetic connections. This is surprising, given the long-standing puzzle of how coral reefs thrive in nutrient-poor waters — a phenomenon commonly called Darwin's paradox because Charles Darwin highlighted this enigma. Tight recycling of energy and nutrients certainly helps⁵, as do the oceanic plankton that are captured by the proverbial "wall of mouths" that the fishes on a reef present⁶.

Graham and colleagues' study adds to our growing appreciation of the importance of long-distance nutritional subsidies for reefs, generated not only by seabirds as documented here, but also by wide-ranging underwater predators such as sharks⁷. Notably, human impacts can disrupt the subsidies in both of these cases.



Figure 1 | Booby chick on a nest above a coral-reef lagoon in the Chagos Archipelago. Graham *et al.*³ report their studies of how non-native rats affect the ecosystems of islands and adjacent coral reefs in the Chagos Archipelago in the Indian Ocean. They find that rat-free islands have substantially more seabirds than do rat-infested islands. Moreover, nitrogen deposits from seabird excrement has a positive effect on nearby coral reefs through nutrient cycling, which generates nitrogen-rich algae that boost the fish population.

This work has immediate practical implications, particularly because reefs are under grave threat around the world. Many of the early losses of coral reefs were due to overfishing⁸, and a scarcity of herbivorous fishes continues to make reefs less resilient^{9,10}. Now, however, one of the major causes for concern over reef survival is the impact of climate change and the ability of reefs to recover from disturbances due to oceanic warming, which is causing mass coral bleaching and death affecting even remote and protected reefs¹¹.

Adding rats to the list of dangers to reefs might seem discouraging. Yet the discovery of the negative impacts of rats on reefs does point directly to a specific strategy that could slow the pace of reef degradation. Rats and other invasive mammals have been successfully eradicated from hundreds of islands¹², with beneficial effects on many terrestrial ecosystems. Graham and colleagues suggest that the same strategy, and others more generally aimed at protecting seabirds, should be a priority for islands associated with coral reefs, helping to buy time while society comes to grips with and, one must hope, slows climate change. In the meantime, scientists will now

be on the lookout for how these rat-infested versus rat-free islands recover from the surely inevitable next coral-bleaching event. ■

Nancy Knowlton is in the Department of Invertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington DC 20013, USA.
e-mail: knowlton@si.edu

1. Harper, G. A. & Bunbury, N. *Glob. Ecol. Conserv.* **3**, 607–627 (2015).
2. Nigro, K. M. *et al.* *Restor. Ecol.* **25**, 1015–1025 (2017).
3. Graham, N. A. J. *et al.* *Nature* **559**, 250–253 (2018).
4. Schill, S. R. *et al.* *PLoS ONE* **10**, e0144199 (2015).
5. de Goeij, J. M. *et al.* *Science* **342**, 108–110 (2013).
6. Hamner, W. M., Colin, P. L. & Hamner, P. P. *Mar. Ecol. Prog. Ser.* **334**, 83–92 (2007).
7. Williams, J. J., Papastamatiou, Y. P., Caselle, J. E., Bradley, D. & Jacoby, D. M. P. *Proc. R. Soc. B* **285**, 20172456 (2018).
8. Pandolfi, J. M. *et al.* *Science* **301**, 955–958 (2003).
9. Jackson, J. B. C., Donovan, M. K., Cramer, K. L. & Lam, V. (eds) *Status and Trends of Caribbean Coral Reefs: 1970–2012* (Glob. Coral Reef Monit. Netw., IUCN, 2014); go.nature.com/2jdpbv
10. Adam, T. C., Burkepile, D. E., Ruttenberg, B. I. & Paddock, M. J. *Mar. Ecol. Prog. Ser.* **520**, 1–20 (2015).
11. Hughes, T. P. *et al.* *Nature* **556**, 492–496 (2018).
12. Russell, J. C. & Holmes, N. D. *Biol. Conserv.* **185**, 1–7 (2015).

China's response to a national land-system sustainability emergency

Brett A. Bryan^{1,2*}, Lei Gao², Yanqiong Ye^{2,3,17}, Xiufeng Sun^{2,4,17}, Jeffery D. Connor^{2,5}, Neville D. Crossman^{2,6}, Mark Stafford-Smith⁷, Jianguo Wu^{8,9}, Chunyang He⁸, Deyong Yu⁸, Zhifeng Liu⁸, Ang Li¹⁰, Qingxu Huang⁸, Hai Ren¹¹, Xiangzheng Deng¹², Hua Zheng¹³, Jianming Niu¹⁴, Guodong Han¹⁵ & Xiangyang Hou¹⁶

China has responded to a national land-system sustainability emergency via an integrated portfolio of large-scale programmes. Here we review 16 sustainability programmes, which invested US\$378.5 billion (in 2015 US\$), covered 623.9 million hectares of land and involved over 500 million people, mostly since 1998. We find overwhelmingly that the interventions improved the sustainability of China's rural land systems, but the impacts are nuanced and adverse outcomes have occurred. We identify some key characteristics of programme success, potential risks to their durability, and future research needs. We suggest directions for China and other nations as they progress towards the Sustainable Development Goals of the United Nations' Agenda 2030.

Exploitation of land, forest, water and nature over thousands of years of human occupation and development had seriously degraded China's environment, impoverished its rural people, and accentuated calamities such as floods, droughts and famine¹. Since the 1950s, political and socio-economic reforms, rapid population rise, industrialization and development, and environmental change accelerated this long-term decline, culminating in a sustainability emergency on a massive scale¹. Multiple natural disasters in the late 1990s spurred the implementation of a portfolio of large-scale policy programmes aimed at mitigating land and water degradation, conserving forests and biodiversity, increasing production from agriculture and forestry, and alleviating rural poverty².

A synthesis of China's recent drive to arrest this long-term decline and ensure sustainability of its land systems is particularly timely given the recent global commitments to the Sustainable Development Goals (SDGs) under the UN's Agenda 2030³. China provides an example of how immense challenges for national-scale SDG implementation⁴ might be addressed. Some studies have provided deep evaluations of China's more well known programmes, particularly the Grain for Green Program and the Natural Forest Conservation Program^{2,5–9}. However, other complementary programmes targeting desertification, grasslands, agriculture and forestry have had comparatively little attention^{10,11}. As a result, the full scale and implications of China's integrated, land-system-wide sustainability response remains under-recognized.

Here, we review 16 major programmes as an integrated sustainability response, with a focus on rural land systems for the period since the establishment of the People's Republic of China in 1949. We quantify the investment and area of actions under these programmes and relate this to the 17 SDGs (Supplementary Methods). We review programme impacts across multiple sustainability indicators. We propose some keys

to success from China's experience, discuss future risks to large-scale sustainability interventions, and suggest future research priorities. China still faces grave environmental concerns resulting from rapid industrialization, urbanization and development, particularly the pollution of its air, water, and soils^{12–15}. However, its recent experience with transformative investment in land-system sustainability can inform how China tackles future challenges and provide invaluable guidance for the rest of the world embarking on a similar journey.

A land-system sustainability emergency

China has been farmed for over 8,000 years¹. Neolithic farmers occupied the North China Plain and the Loess Plateau (dryland millet) and the Yangtze River valley (wet rice). Over time, forests were progressively cleared for agriculture and exploited for energy, food, medicines and materials; cropping intruded into northern grasslands; and rivers were redirected for irrigation¹. Over the first 1,400 years AD, nomadic pastoralism expanded in the north and increasingly productive wet-rice farming spread further south, supporting a fairly stable population of 40–60 million¹⁶. From 1400 to 1750 China's population increased 3–4-fold to 177 million (0.24% per year)¹⁶ supported by agricultural expansion, commerce, markets and trade, with forest cover at around 24.2%¹⁷. By 1800¹, humans occupied most of China, and population growth accelerated (0.59% per year), reaching around 580 million by 1949¹⁶, alongside an emerging environmental crisis. Increasing demand for natural resources, particularly agricultural and forest products, met by inefficient and unsustainable farming, grazing and logging practices led to widespread land degradation including erosion, sedimentation, flooding, water logging, salinization and desertification¹. Bears, tigers, leopards, elephants and many other species neared extinction¹⁸. Soil-nutrient depletion constrained agricultural productivity. Forest cover

¹Centre for Integrative Ecology, Deakin University, Geelong, Victoria, Australia. ²CSIRO, Waite Campus, Adelaide, South Australia, Australia. ³College of Natural Resources and Environment, South China Agricultural University, Guangzhou, China. ⁴College of Horticulture and Landscape Architecture, Southwest University, Chongqing, China. ⁵School of Commerce, City West Campus, University of South Australia, Adelaide, South Australia, Australia. ⁶School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, Australia. ⁷CSIRO, Black Mountain, Canberra, Australian Capital Territory, Australia. ⁸Center for Human-Environment System Sustainability (CHESS), State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing, China. ⁹School of Life Sciences and School of Sustainability, Arizona State University, Tempe, AZ, USA. ¹⁰State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing, China. ¹¹Key Laboratory of Vegetation Restoration and Management of Degraded Ecosystems, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China. ¹²Center for Chinese Agricultural Policy, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. ¹³State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China. ¹⁴School of Ecology and Environment, Inner Mongolia University, Hohhot, China. ¹⁵College of Ecology and Environmental Science, Inner Mongolia Agricultural University, Hohhot, China. ¹⁶National Forage Improvement Center, Key Laboratory of Grassland Resources and Utilization of Ministry of Agriculture, Institute of Grassland Research, Chinese Academy of Agricultural Sciences, Hohhot, China. ¹⁷These authors contributed equally: Yanqiong Ye, Xiufeng Sun. *e-mail: b.bryan@deakin.edu.au

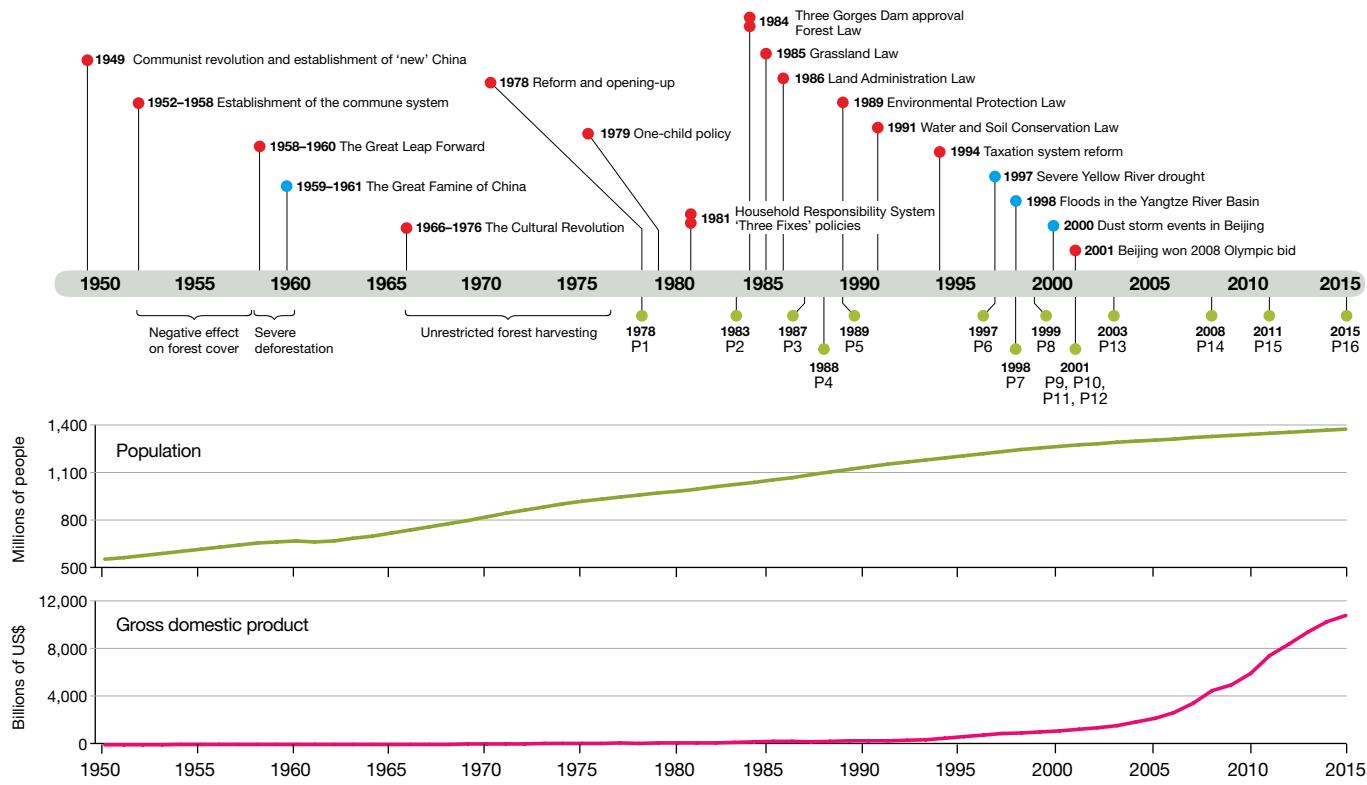


Fig. 1 | Timeline of events. Chronology of key political/socio-economic/policy events (red markers), human-environmental disasters (blue markers), and the 16 major sustainability programme responses (green

markers) in China from 1949 to 2015. Also included are the trajectories of population and economic development (gross domestic product or GDP).

was around 11.4%¹⁷ or less¹ and timber was scarce. Rural poverty, social unrest and famine were commonplace¹.

In 1949, the new People's Republic of China inherited a seriously degraded natural environment¹⁸ and in the Mao era through to 1978, China's land, rivers and forests were seen as forces to be controlled and resources to be exploited to support food security, economic development and industrialization^{1,5}. Population growth rates increased to 1.36% per year under the new government¹⁶. In the early 1950s, anticipation of collective ownership policies led to a rush to clear land² (Fig. 1). Deforestation intensified under the Great Leap Forward as state forestry bureaus and commune enterprises decimated the meagre forest resources for timber and fuel, in particular for steel-making⁵. Droughts and floods compounded multiple governance failures during the 1950s, which included poor planning, redirection of agricultural labour to industry, and the depletion and inadequate distribution of scarce resources including food. Catastrophic outcomes ensued, most notably the Great Famine of China from 1959–1961 which claimed the lives of 20–45 million people^{1,19}. Rivers, particularly the Huai and Yellow, were dammed and dyked for irrigation and flood management¹. Pursuit of self-sufficiency in grain led to an increase in agricultural production, mostly via farmland expansion (from 80 million hectares (Mha) to around 120–130 Mha in the first 30 years of the People's Republic)^{1,20,21}. Forest cover dropped to 8.7% in the 1960s¹⁷. Clearance of the northeastern and southwestern forests, exploitation of the northern and northwestern grasslands, combined with inappropriate farming practices in marginal environments (such as steep slopes, flood plains and erodible soils) accelerated land, water and ecosystem degradation¹. During the period of the Cultural Revolution from 1966–1976, environmental degradation was exacerbated by diminished government capacity to manage natural resources⁵.

The 1978 Reform and Opening Policy introduced a range of free-market and privatization measures to promote development. The Household Responsibility System, and the re-privatization of land from communes to households led to greatly increased agricultural productivity and expansion, but often via unsustainable practices on marginal land, particularly the overgrazing of grasslands²². The One-Child

Policy (1979) was introduced to control population growth and reduce demand on natural resources. Improving domestic capacity for chemical fertilizer production²³ gradually led to agricultural surpluses¹. Rapid economic and social development followed, along with industrialization and urbanization, and an expanding, increasingly wealthy, and technologically advancing population. While this reduced people's direct dependence on land for livelihoods, it also fuelled consumption, in particular demand for forest and agricultural products¹². The opening of commercial timber markets in the early 1980s accelerated forest harvesting by the rural poor at rates far outpacing natural regeneration and reforestation. Reforms in the 1980s included the reinstatement of the Ministry of Forestry and the enactment of key laws (Fig. 1), but deforestation continued into the 1990s^{5,20,24} and beyond²⁵.

Extreme land degradation ensued throughout the 1970s, 1980s and 1990s. While plantation efforts reversed the decline in total forest cover during this period^{1,17,20} (with the definition of 'forest' broadly including a range of vegetation types such as natural forest, orchards, rubber, timber and shelter plantations^{26,27}), natural forest decline continued from 102.0 Mha in 1949 to 98.2 Mha in 1975 and 66.7 Mha in 1993, with much of the remaining forests degraded and unproductive². Soil erosion of about 5 billion tonnes annually affected 360 Mha²⁸, including 75 Mha of the Yangtze and Yellow river basins, and caused major water quality, sedimentation, and flooding problems². In the worst-affected areas of the Loess Plateau, erosion rates reached $100 \text{ t ha}^{-1} \text{ yr}^{-1}$ and the sediment load²⁹ of the Yellow River—the world's muddiest—reached 1.8 billion t yr⁻¹. Growth in agricultural production was also down, jeopardizing food security³⁰. Over 54 Mha (40%) of existing arable land was degraded, and 70% of it was of lower productivity³¹. In northern China, rangelands were degraded and desertification intensified³², reaching 267.4 Mha in 1999³³. Causes were both human, particularly through the expansion and intensification of livestock production³⁴, and environmental, with the naturally sandy soils and the dry, mid-latitude, continental climate becoming drier and windier^{35,36}. As poor households with low education and few prospects for off-farm employment overexploited farmlands, pastures, and forests to provide food and

Box 1

Programme aims

This is a summary of the planned timeframe, aims, and objectives of the 16 major Chinese sustainability programmes assessed here. Detailed descriptions and sources are provided in Supplementary Tables 1–16.

P1 Shelterbelt Development Program—Three North. 1978–2050. Control the expansion of sandy/desertified land, and mitigate wind erosion of sand/soil and dust storms in northern China via forest plantation, mountain closure, and sandy area regeneration.

P2 Soil and Water Conservation Program—National. 1983–2017. Control soil erosion; improve farmers' livelihoods; and improve agricultural production, ecology, and the environment by combining prevention, protection, control, repair and ecological regeneration, and utilizing appropriate scientific, engineering, plantation and cultivation measures.

P3 Shelterbelt Development Program—Five Regions. 1987–2020. Arrest environmental deterioration in the Yangtze River, Pearl River, their coastal areas, the Plain, and the Taihang Mountains via artificial plantation, mountain closure, aerial seeding, improving low-yielding forest and establishing shelterbelts.

P4 Comprehensive Agricultural Development Program. 1988–2020. Raise rural quality of life, incomes and food security through land reform, land management, ecological construction, agricultural infrastructure and industry development, and production/efficiency gains using science and technology.

P5 Soil and Water Conservation Program—Yangtze. 1989–indefinite. Reduce sedimentation and improve the health of the Yangtze River, ensure the safe operation of the Three Gorges Reservoir, and enhance regional economic and social development by controlling soil erosion in the upper reaches.

P6 National Land Consolidation Program. 1997–2020. Increase the area of cultivated land and revenues via consolidation (reorganizing and merging fragmented and underused land), reclamation, constructing high-quality cropland, and improving land use and management.

P7 Natural Forest Conservation Program. 1998–2020. Halt logging/deforestation and protect natural forests for ecological/carbon benefits via mountain closure, aerial seeding and artificial planting. Create new business opportunities for traditional forest enterprises; create forest management jobs and relocate redundant forestry workers.

P8 Grain for Green Program. 1999–2020. Prevent soil erosion, mitigate flooding, store carbon, and improve livelihoods by increasing forest and grassland cover on cropped hillslopes and converting cropland, barren hills and wasteland to forest.

P9 Fast-growing and High-yielding Timber Program. 2001–2015. Remedy the decline in timber supply and meet domestic demand for forest resources without affecting natural forests via the establishment of fast-growing and high-yielding timber plantations.

P10 Forest Ecosystem Compensation Fund. 2001–2016. Conserve natural forests and protect species and ecosystems via restoration, protection, and management of forests that have important ecological, biodiversity conservation, and sustainable economic and social value.

P11 Sandification Control Program—Beijing/Tianjin. 2001–2022. Reduce desertification and dust storms, and improve the environment in the Beijing/Tianjin area via reforestation, grassland management, and water conservation, relocating affected people and establishing basic governance of desertified lands.

P12 Wildlife Conservation and Nature Protection Program. 2001–2050. Conserve key wild animal and plant species and natural ecosystems by expanding the number and area of nature reserves, and promoting sustainable development.

P13 Partnership to Combat Land Degradation. 2003–2023. Improve management of land and water resources, reduce poverty, protect biodiversity, and combat climate change in western China by bringing agencies together in partnership to work synergistically.

P14 Rocky Desertification Treatment Program. 2008–2020. Curb rocky/karst desertification, improve the environment, and increase incomes by protecting and establishing vegetation, promoting sustainable land-use, farmland construction, water conservation and relocating poor people.

P15 Grassland Ecological Protection Program. 2011–2020. Mitigate grassland degradation by grazing prohibition and enhancing grassland vegetation coverage/biomass. Increase herder incomes by promoting the sustainable development of pastoral areas.

P16 Cultivated Land Quality Program. 2015–2030. Enhance food security and the quality, safety and ecological sustainability of agricultural production by addressing soil acidification, salinization, nutrient imbalances, pollution, biota, fertility and shallow topsoil.

income, the productivity of these increasingly degraded environments was further diminished, creating a poverty–environmental degradation trap³⁷. The economic impacts of land degradation and desertification (0.7%–1.4% of GDP in 1999) impeded rural development³⁸.

From 1978–1997, the Chinese government launched six programmes to address the parlous state of its rural land systems (P1–P6; Fig. 1, Box 1, Supplementary Tables 1–6). However, modest initial efforts failed to arrest the deteriorating trend and, in the late 1990s, China suffered a series of natural disasters widely believed to be caused by unsustainable land management⁵. In 1997, drought along the Yellow River, greatly amplified by water over-extraction, desiccated 700 km of the waterway for 8 months³⁹. The 1998 Yangtze River floods—among China's most devastating, killing 3,600 people, inundating 5 Mha of cropland, and costing US\$36 billion—were driven by El Niño but exacerbated by upland deforestation⁴⁰. In the spring of 2000, a sequence of 12 severe dust storms afflicted northern China (Fig. 1), with seven blanketing Beijing within just one month⁶. The storms were attributed to overgrazing and desertification⁴¹ and cost US\$2.2 billion⁴².

The response

In response to this sustainability emergency, China dramatically ramped up investment in its six existing programmes and launched

seven major new programmes from 1998–2003 (P7–P13), complemented in later years by three additional programmes (P14–P16; Fig. 1). Major foci were to reduce erosion, sedimentation and flooding in the Yangtze and Yellow rivers; conserve forest in the northeast; mitigate desertification and dust storms in the dry north and rocky south; and increase agricultural productivity in central and eastern China. Environmental objectives were typically complemented by strong socio-economic objectives such as poverty reduction, rural economic development, and national food security (Box 1, Supplementary Tables 1–16).

Investment in the 16 major sustainability programmes from 1978–2015 totalled US\$378.5 billion (in 2015 US dollars). US\$351.6 billion was invested from 1998–2015, with total annual investment increasing steadily as China's economy grew, from US\$3.52 billion in 1998 (about 0.34% of GDP) to US\$40.6 billion in 2015 (about 0.37% of GDP) (Figs. 1 and 2a and b). Programmes addressed a combined area of 623.9 Mha, or 65% of China's land area. Total area increases from 1998–2015 averaged 32 Mha yr⁻¹ (not including the 252.1 Mha added in 2011 by the P15 Grassland Ecological Protection Program) (Fig. 2d and e). For reference, China's investment far exceeds other globally important national sustainability programmes such as the US Conservation Reserve Program, where rental payments totalled \$46.2 billion from

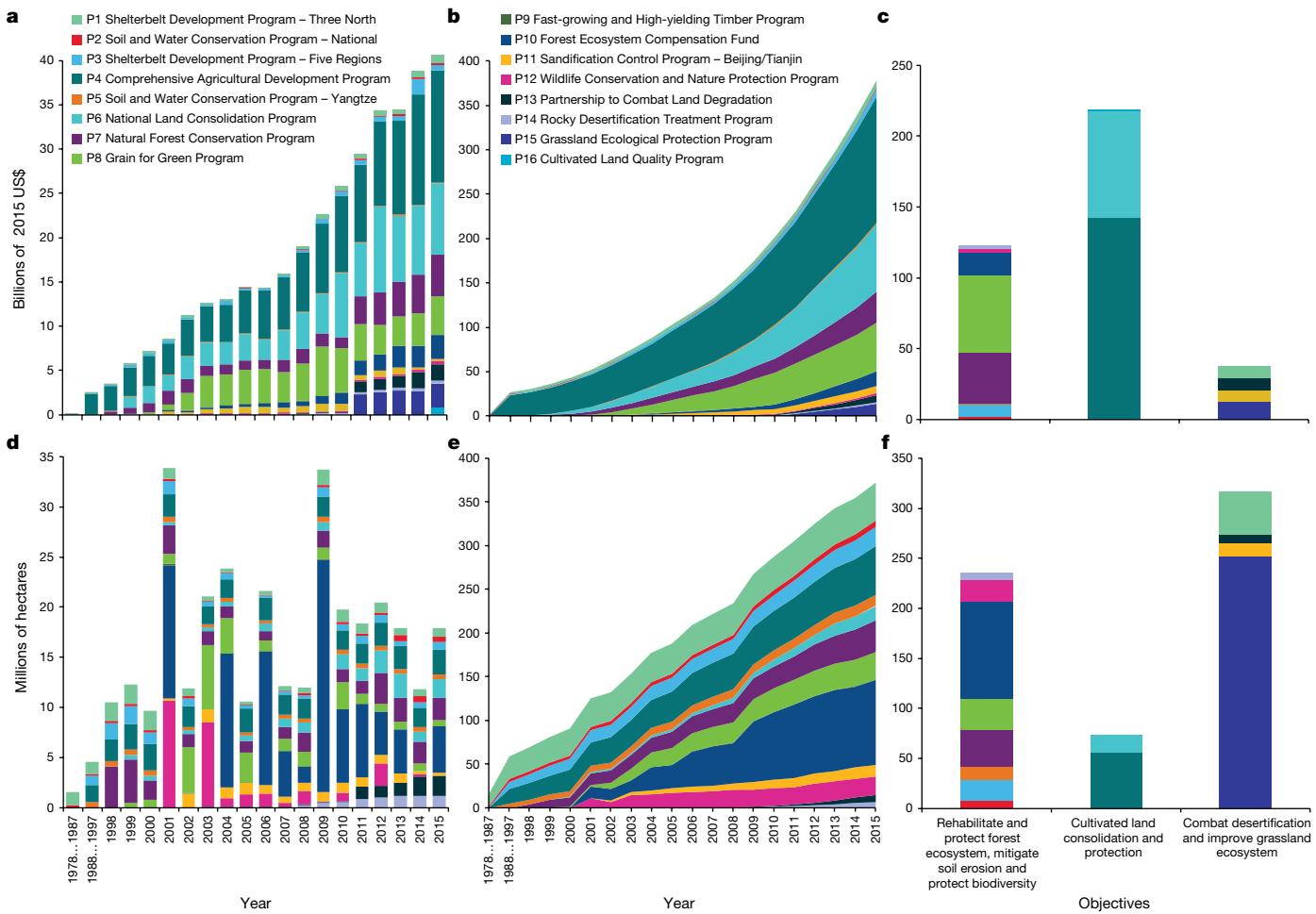


Fig. 2 | Investment and area of interventions. Value of investment (a, annual; b, cumulative; c, by objective) and area (d, annual; e, cumulative; f, by objective) addressed over time for the 16 major sustainability programmes reviewed. Values for the decades 1978–1987 and 1988–1997 are annual averages in a and d, and summed in b and e. For readability, area graphs d and e do not include the P15 Grassland Ecological

Protection Program. Data includes area and investment values for nature reserves established from 2001, when the Wildlife Conservation and Nature Reserve Program became a priority programme. 2015 values are interpolated values for some programmes where official data have not yet been published. Data sources and methods are detailed in Supplementary Methods, and the full dataset is provided in Supplementary Data 1 and 2.

1987–2016 (around 0.0094% of the USA's GDP in 2014), and averaged 12.6 Mha annually⁴³.

Multiple programmes operated within each province (Supplementary Tables 19 and 20). Agricultural production and cultivated land protection dominated investment but covered the least area (Fig. 2c and f) owing to the higher cost of intensifying and expanding agriculture and increasing productivity. Forest ecosystem protection, reforestation, alleviating soil erosion, and protecting biodiversity drew the second-highest investment and area of actions. Combating desertification and improving dryland ecosystems involved the least investment but covered the greatest area owing to the lower cost of extensive measures such as grazing prohibition and grassland restoration. Across the portfolio, an array of interventions (Supplementary Fig. 1) were implemented to address the multiple socio-economic and environmental objectives (Box 1), with each programme typically including multiple interventions (Supplementary Data 3). Implementation area varied markedly between actions and over time (Supplementary Fig. 1, Supplementary Table 21).

Programme governance was led primarily by the central government (Supplementary Tables 1–16; Fig. 3), which also provided most of the funding, supported by partnering and co-investment from provincial/local governments, enterprises and individuals. With the help of research agencies, the central government designed the sustainability programmes, set high-level objectives, and delegated responsibility to relevant ministries, commissions and administrations

(such as the State Forestry Administration). These agencies planned programme scope and priorities, and coordinated implementation, allocating tasks to provincial (including autonomous regions and independent municipalities) government departments. Overseen by their respective governments, provincial/local government departments refined and adapted programme plans based on regional/local conditions and priorities; and developed and implemented projects, managed funding, and supervised and inspected sustainability interventions. Coordination and communication between governments/departments was directed by governance, guidance/supervision and reporting processes. All levels of provincial/local government and their departments raised the awareness and enthusiasm of farmers, herders and enterprises, and mobilized and organized them to implement large-scale sustainability interventions. Public opinion agencies were used to promote sustainability programmes, report progress to the people, and report local society's attitudes and responses to the programmes, which also fed back to government. Research agencies identified potential issues, assessed programme effectiveness, verified implementation, and provided scientific and technological support to implementation agents (such as farmers and enterprises). Monitoring and quality assurance of sustainability interventions usually involved self-appraisal; inspection at the local, provincial and national levels; and verification against accepted standards. Under-performance incurred penalties including withheld payments (Supplementary Tables 1–16; Fig. 3).

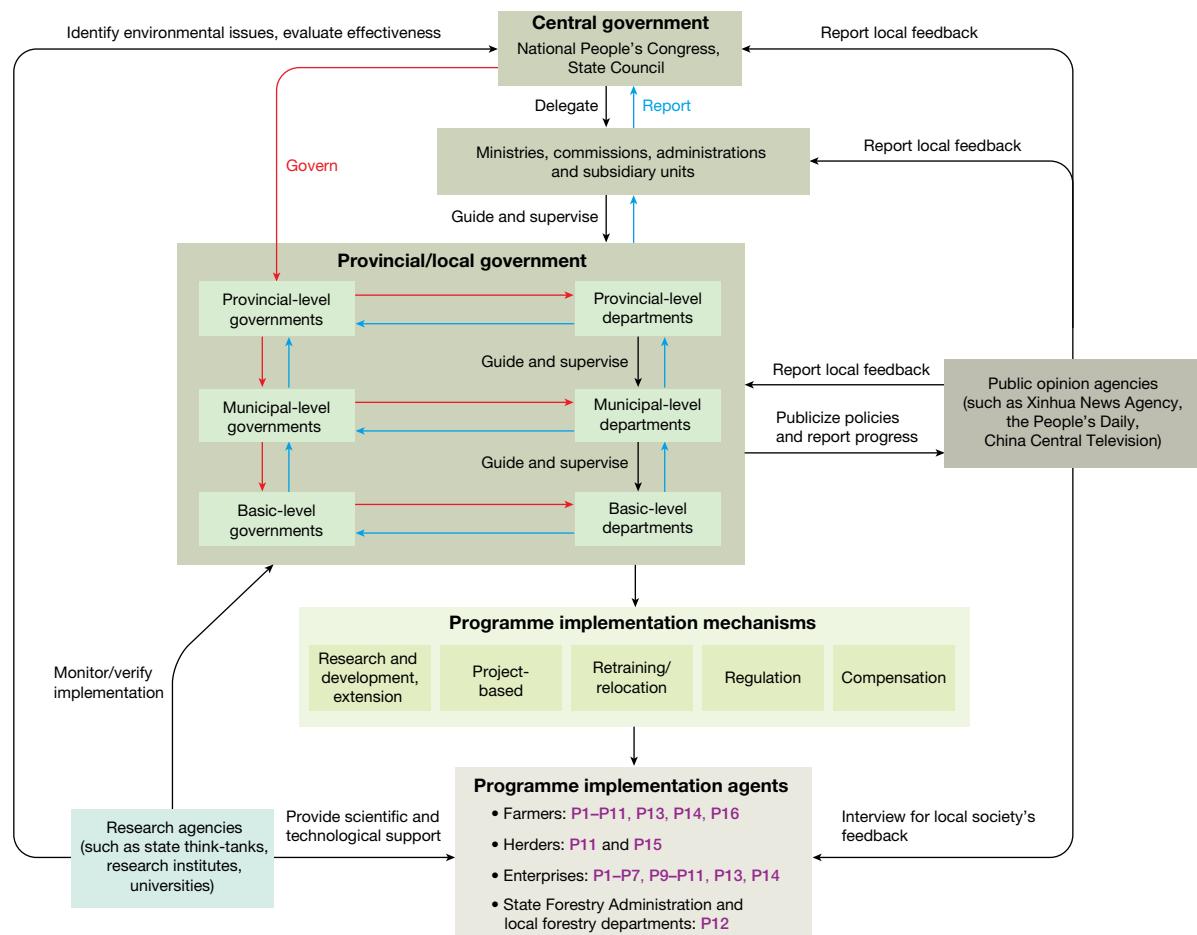


Fig. 3 | Governance, administration and implementation of China's sustainability programmes. Entities are shown in shaded boxes and arrows indicate specific roles, processes and information flows

(red arrows, 'Govern'; blue arrows, 'Report'). Examples (purple font) are included to illustrate the programmes engaging each type of implementation agent.

Most programmes employed a variety of implementation mechanisms (such as regulation and compensation) to engage implementation agents (Supplementary Tables 1–16; Fig. 3). Motivating voluntary participation was a common objective, as farmers, landholders and local communities often provided most of the labour for on-ground works. Compensation via direct payments to people for the provision of ecosystem services was widely used to simultaneously reduce poverty and incentivize landholders to change land use and management⁴⁴. While property rights reforms and de-collectivization policies provided an essential foundation for sustainable land management, additional regulatory support was necessary to stop over-exploitation³⁷. Regulatory controls such as livestock prohibition and timber quotas were used to end unsustainable practices in high-priority areas⁷. Where livelihoods were affected by resource-use regulation, compensation was typically provided as well as retraining and re-employment opportunities such as the conversion of loggers into forest managers^{2,8,24}. In the areas worst affected by environmental degradation and poverty, sustainability interventions were preceded by the relocation of local people to relieve environmental pressure⁴⁵. Despite notable exceptions⁴⁶, this process of 'ecological migration' was largely voluntary, with households offered education/training, modern housing and land, employment, and financial assistance to move to less-sensitive areas^{45,47}. Programmes also engaged enterprises in large-scale project-based interventions and infrastructure including timber plantations; restoration supply chains (such as plant nurseries, seed collection); water supply and drainage management; and transport, energy and telecommunication networks^{6,19}. Research, development, application, and extension of technologies such as trait-based plant species selection for erosion control⁴⁸, straw-checkerboard desertification

control⁴⁹, and the biological control of forest pests (such as the fall webworm or mulberry moth, *Hyphantria cunea*)⁵⁰ were also important components of many programmes.

Hundreds of millions of people have participated in and been affected by the programmes in various ways (Supplementary Tables 1–16). A vast labour force implemented on-ground sustainability interventions. For example, over 124 million farmers (>32 million rural households) across 2,279 counties and 25 provinces have directly benefited from the Grain for Green Program^{5,8}. About 170 million people have been involved in the Shelterbelt Development Program—Three North across 13 provinces and 600 counties⁵¹. The Soil and Water Conservation Program—Yangtze has mobilized more than 2.1 billion working days of labour towards sustainability interventions⁵². Over 7.7 million people have been relocated from ecologically degraded regions to reduce poverty and environmental pressure⁵³. One of the largest examples is the relocation of around one million Hui ethnic minority farmers and herders from the highly degraded central drylands and southern mountains of Ningxia to northern Yellow River irrigation areas⁵⁴. The Natural Forest Conservation Program affected 1.2 million logging and processing workers, most of whom have been transferred (to plantation and forest management activities or to other sectors), retired, or laid off^{2,6,8}. Under the Comprehensive Agricultural Development Program, 1.236 million people underwent technical training⁵⁵.

Programme impacts

China has undertaken unprecedented investment in sustainability since 1998, addressing vast areas of land. Below, we synthesize current understanding of national-scale impacts on key sustainability indicators and illustrate impacts via regional/local case studies.

Forests and grasslands

China's forest cover transitioned in recent decades, turning from net loss to gain^{17,20}, and reaching 22.2% national coverage in 2015⁵⁶. Programmes such as Grain for Green and the Shelterbelt Development Program—Three North have greatly increased forest cover through reforestation and afforestation (planting forests where forests had never before grown) of 60.15 Mha from 1998–2014 (Supplementary Fig. 1, Supplementary Data 3). Vegetation cover nearly doubled on the Loess Plateau from 1999–2013, increasing from 31.6% to 59.6% largely due to Grain-for-Green conversion of agricultural land on slopes exceeding 15°⁵⁷. Similarly, multi-programme afforestation of grasslands in Xinjiang in China's arid north-west increased forest cover by 68% from 2000–2009⁵⁸.

The vast majority of new forests were 'protection forests' (47.156 Mha from 1998–2014) for stabilizing degraded land, 'economic forests' (6.295 Mha) such as fruit orchards and rubber plantations, and timber forests (6.061 Mha) (Supplementary Fig. 1, Supplementary Data 3). However, natural forest ecosystems have also benefited from the 133.477 Mha of forest protection and management interventions such as mountain closures and logging bans, and 21.695 Mha of nature reserves established from 2001–2014 (adding to the 113.629 Mha established from 1956–2000), which have slowed deforestation, promoted regeneration and enhanced ecological condition^{25,59}. The Natural Forest Conservation Program for example, has been associated with large gains in forest cover from 2000–2010⁶⁰, and in target provinces has reduced annual deforestation rates to 0.62% (3.3 times lower than in other provinces)²⁵.

Grassland ecosystems in northern and western China have responded to large-scale restoration⁶¹ and grazing exclusion⁶² across 260.276 Mha (Supplementary Fig. 1, Supplementary Data 3). Even before the Grassland Ecological Protection Program boosted grassland conservation in 2011, Inner Mongolia's grassland increased by 7.799 Mha via conversion from desertified land and cropland, at an average net primary productivity (NPP) gain of 29,433 gigagrams of carbon per year (GgC yr⁻¹) from 2001–2009. Roughly 80% of these gains were attributed to programme restoration and de-stocking and 20% to climate change⁶³. From 2005–2012, in the Tibetan Plateau Yellow River headwaters, nomadic herder resettlement, protection from livestock, and restoration interventions under multiple programmes have slowed and reversed alpine meadow degradation⁶⁴.

Desertification and dust storms

The desertification trend, dominant since the 1950s in China's arid northwest and semi-arid north and northeast regions, has also reversed over the past two decades^{35,65,66}. National monitoring reported a decrease in desertified land in China, declining from a peak of 267.4 Mha in 1999 to 261.1 Mha in 2014^{33,38}. Although primarily controlled by climate^{65,66}, in particular reductions in wind speed and increased spring rainfall³⁵, the greening trend has been enhanced by sustainability interventions in target areas⁶⁶.

In China's north, greenness and NPP have recently increased overall, indicating a reversal of the desertifying trend, but this varied over space and time⁶⁷. Some areas, particularly in the northeast (Inner Mongolia, Heilongjiang) and northwest (Xingjiang), continued to degrade; while other areas (such as Qinghai, Gansu, Shaanxi and Shanxi) greened^{59,66,67}. Three North greening at 0.86%–1.12% yr⁻¹ from 2000–2013 was faster than the national average and faster than the period 1982–2000 (0.28%–0.38% yr⁻¹)⁶⁸. While climate had the strongest influence (74%) on NPP changes along with other natural and anthropogenic factors (23%), sustainability programmes had a discernible impact (3%)⁶⁷. Barren and sparsely vegetated land in the Three North region decreased by 7 Mha to 178.65 Mha from 2001 to 2010⁶⁷. In the Beijing/Tianjin Sand Source Region, degradation increased over 6.9%–10.8% of land and decreased over 3.8%–7.0% of land from 2000–2010, with these changes spatially heterogeneous and climate-driven⁶⁹. Dust storms in northern China have also declined since the 1950s^{33,66} and while restoration and afforestation may have

contributed to this, lessening wind speed and reduced frequency of windy days are the more likely reasons⁶⁶.

Soil and water

An overall decrease in soil erosion of 12.9% has been identified nationally from 2000 to 2010⁹. In 11 of China's major river systems, soil erosion decreased by 45.4% on average from 2003–2007 compared to the period 1998–2002, including 58.8% and 27% declines in the Yangtze and Yellow river basins, respectively, associated with large-scale Grain-for-Green restoration⁷⁰. In the Loess Plateau, large-scale restoration and afforestation of cropland and barren land has reduced soil erosion to historically low levels⁵⁷. Illustratively, in the Zuli basin, modelled estimates of a net 25.7% ± 8.5% reduction in soil erosion from 1999 to 2006 included a 38.8% restoration-induced decrease, counteracted by a 13.1% ± 4.3% rainfall-induced increase⁷¹. Integrated analysis of satellite imagery, restoration statistics and sediment yield monitoring confirmed that large-scale conversion of cropland to forest had a positive impact on greening and mitigation of soil erosion across southern China⁷². Complementary consolidation of farmland in valley bottoms has relieved agricultural development pressure on sloping land and reduced soil erosion by a further 10%⁷³.

China's sustainability programmes have achieved a range of water management objectives such as improving water quality and reducing river sedimentation, soil water retention and flood mitigation, and water conservation and supply (Supplementary Tables 1–16). Nationwide, improvements in flood mitigation and water retention of 12.7% and 3.6%, respectively, occurred from 2000–2010⁹. Substantial improvements in surface water quality and sedimentation have also resulted from a combination of source-control interventions (such as restoration and land consolidation) and in-stream measures (such as silt check dams and reservoirs)²⁷. Since 1950, the sediment load of the Yellow River has fallen by nearly 90%⁷⁴. From 1980 to 1999, cropland terracing and dams/reservoirs reduced sediment load by 33% and 21%, respectively, and large-scale afforestation further reduced sediment loads by 26% from 2000 to 2010⁷⁴. Similarly, Yangtze River sediment load is down 71% since the mid-1900s. Soil and water conservation measures explain 6%–10% of this change, with the Three Gorges Dam accounting for 31%–65% and other dams accounting for 10%–57%⁷⁵.

A major consequence of large-scale afforestation has been the impact on water resources via increased evapotranspiration and decreased run-off, stream flow and groundwater, especially in drylands^{76,77}. For example, field studies have found soil water depletion (0–1 m depth) of 14.5%–42.0% under Grain-for-Green restoration⁷⁶, and afforestation has lowered local groundwater tables between 0.5 and 3.0 m in China's drylands⁷⁸. Soil and groundwater depletion has reduced the survival rates of afforestation to 7%–34% in some areas^{78,79}, limiting its effectiveness in desertification control^{35,77}. Scaled up, the water impacts resulting from the large-scale mismatch of species' water requirements with water availability may be important given northern China's water scarcity^{23,77,80}. Afforestation in the Loess Plateau, for example, is approaching the sustainable water resource-use limits, affecting human use, and threatening local/regional food security⁸¹. However, over the entire Three North region, with afforestation covering just 7.8% from 1998–2014 (Supplementary Data 3), climatic variability/change is likely to have had a far greater water impact³⁶. Nonetheless, dryland sustainability programmes need to ensure that the water requirements of species used in large-scale restoration are compatible with local/regional environmental water availability.

Biodiversity

A slight nationwide decline (–3.1%) in ecological habitat from 2000 to 2010 has been reported⁹. However, programmes such as the Natural Forest Conservation Program, which implemented 22.645 Mha of mountain closure and forest tending from 1998 to 2014 (Supplementary Fig. 1, Supplementary Table 21), have substantially slowed the decline in China's natural biodiversity^{8,60}. China's nature reserve system also expanded rapidly after 1992 and was consolidated under the Wildlife

Conservation and Nature Reserve Program in 2001. By 2014, it comprised 2,729 reserves and covered around 15.1% of China's territory⁸² (Supplementary Fig. 1, Supplementary Data 3). The Forest Ecosystem Compensation Fund has also been implemented over a large area (97.334 Mha) but its effectiveness has not been evaluated. Ex situ conservation measures such as breeding/relocation programmes, botanical/zoo logical gardens, and gene/germplasm banks have complemented the in situ measures above⁷. Local success stories include the Qinling Mountains biodiversity hotspot with its simultaneous forest protection and recovery, economic development and giant panda population increases⁸³.

However, there is plenty of room for improvement in biodiversity outcomes across China's sustainability portfolio. Whereas it covers 17.9% of the habitat area of threatened mammals, China's reserve system represents only 8.5% of reptile habitat and poorly represents ecosystem services⁸². China's adoption of the broad definition of 'forest' ignores differences in habitat function and biodiversity values of different forest types⁸⁴. Reports of great forest area increases belie the predominance of non-native, single-species plantations, and the afforestation of vast areas that had never supported forest before, with their attendant biodiversity impacts²⁶. In some areas, unintended consequences for biodiversity have resulted from policy incentives motivating landholders to first fell natural forest for sale, then establish programme-funded plantations on the newly-cleared land^{26,84}. In addition, fast-growing pioneer tree and forage crop species such as pines (*Pinus* spp.), poplars (*Populus* spp.), and willows (*Salix* spp.) widely used to remediate infertile and eroding soils, have also invaded natural ecosystems⁸⁵. Impacts on animal populations were also evident, as demonstrated in Sichuan where, compared to natural forest, compositionally simple reforestation was 17%–61% and 49%–91% lower in bird and bee diversity, respectively⁸⁶. Prioritizing the restoration of natural ecosystems over monoculture plantations in key areas could substantially improve biodiversity outcomes at low cost⁸⁶.

Agriculture and food

From 1985 to 2007, China's agricultural outputs grew 5.1% per year on average following the 1978 reforms⁸⁷. County-level crop production data revealed a near-doubling of cereal production from 1980 to 2010, but improvements varied by farming system and were limited by environmental constraints and climatic variability³⁰. Rice yields increased over 12.3 Mha (41.8% of China's rice-growing area), but stagnated over 14.7 Mha (50%); wheat yields increased over 13.8 Mha (58.2%), but stagnated over 3.8 Mha (15.8%); and maize yields increased over 5.3 Mha (17.7%), but stagnated over 16.3 Mha (54%)³⁰. From 2005 to 2010, meat production increased by 14%, egg production by 28%, and milk by 38%⁸⁸. In a national survey from 2005–2009, increases in intermediate inputs (seed, fertilizer, irrigation and machinery) accounted for 44.46% of grain production growth, cropped area 18.16%, and Total Factor Productivity growth 17.30%⁸⁹. Hunger has largely disappeared in China following the substantial increase in per-capita food production⁹⁰. However, intensification and decreased nutrient-use efficiency in both crop⁹¹ and livestock⁹² systems have led to widespread decline in lake, river and coastal water quality⁹³.

The Comprehensive Agricultural Development Program reported 28.096 Mha of low-and medium-yield cropland improvement and 4.114 Mha of high-yield cropland demonstration from 1998 to 2014. The National Land Consolidation Program reported 14.978 Mha of agricultural land consolidation, development and reclamation, and 1.968 Mha of small watershed improvement was reported under multiple programmes (Supplementary Fig. 1, Supplementary Table 21, Supplementary Data 3). 29.5% of land parcels that underwent consolidation/reclamation showed improved productivity, and this increased over time⁹⁴. While substantial areas of farmland have been lost to urbanization and ecological restoration, agricultural development and land consolidation programmes have more than offset these area losses⁹⁵. However, productivity suffered because new farmland was often of lower quality than displaced areas⁹⁶.

Sustainability programmes have contributed to the growth in China's agricultural production in many ways. For example, productivity growth in Wuqi County of 15.8% from 1998 to 2004 was attributed to technical improvements resulting from Grain-for-Green extension services and diffusion of technical knowledge⁹⁷. Technological advances such as semi-dwarf rice and wheat cultivars, heterosis in rice and maize, and mitigation of soil salinization have also boosted productivity⁹⁰. The impact of large-scale environmental restoration on agricultural production, food prices and imports has been minimized via the targeting of steep, lower-productivity land⁹⁸. Payments have relaxed farm household liquidity constraints, enabling intensification of agricultural production⁹⁹ and increased cropped areas¹⁰⁰. Other policies such as agricultural and land tenure reforms, land-use regulation, and agricultural subsidies have also strongly affected agricultural production in China⁸⁷ but further research is needed to tease out their relative influence.

Society and economy

China's sustainability programmes have generally increased incomes and reduced poverty, but the effects have varied^{5,101,102}. From 1995 to 2004, of six priority forestry programmes analysed, one was found to have strongly positive short-term impacts on rural household incomes (P6), three had positive impacts (P1, P3, P7), one was negligible (P11), and one was weakly negative (P12)¹⁰³. Income effects became more positive over time however, as households adjusted and as market and environmental conditions improved¹⁰¹.

Best understood are the socio-economic impacts of the Grain for Green Program and the Natural Forest Conservation Program. While Grain-for-Green payments have had a small direct effect on household incomes, the greater effect has typically been indirect, where relaxed liquidity constraints have freed up farm household labour for higher-paying off-farm employment¹⁰⁴. The more land enrolled, the greater the effect¹⁰⁵. In a large-sample longitudinal study in Shaanxi and Sichuan from 1999 to 2008, a 250% increase in household incomes was found, mainly driven by off-farm employment⁹⁹. The socio-economic effects of the Natural Forest Conservation Program have been more mixed. In northern Shaanxi for example, 34.9%, 47.0% and 59.8% of farmers, pastoralists and forest workers, respectively, reported adverse impacts on their livelihoods due to logging and grazing bans, inadequate compensation and cost-sharing expectations. 23.5% of former forest workers remained unemployed after 6 years and livestock incomes were affected by the higher costs of barn-raising animals¹⁰⁶. In the Wolong Nature Reserve, the Natural Forest Conservation Program had both positive (for example, less labour required to collect fuel-wood) and negative (for example, crop raiding by wildlife) effects on households, and while household incomes quadrupled from 1998 to 2007, increases were largely due to engagement in agriculture, off-farm employment and the tourism industry⁴⁴.

Ecological migration has lifted many out of abject poverty and dire environmental conditions, particularly in China's north and west^{53,107–109}. Benefits include improvements in source-area environments; incomes and living standards; access to housing, utilities, healthcare, transport, and education; safety from floods and landslides; and social engagement and cohesion^{47,108}. While initiatives such as the Massive Southern Shaanxi Migration Program and the Ningxia Hui migration have reported overall success^{47,108}, in many cases, ecological migration has also placed pressure on the receiving environment and caused a range of socio-economic problems^{45,109}. New houses and land endowments have been too small to meet needs^{54,108}; and migrants, often poorly educated, can be unsuited to off-farm employment¹¹⁰. Households also suffered from increased cost of living, loss of social networks⁴⁷, and cultural disruption such as the relocation of nomadic tribes and conversion to modern, sedentary urban life^{46,107,109}.

Addressing the SDGs

Despite the adoption of the SDGs only at the end of our study period (2015), it is instructive to visualize China's sustainability investment in

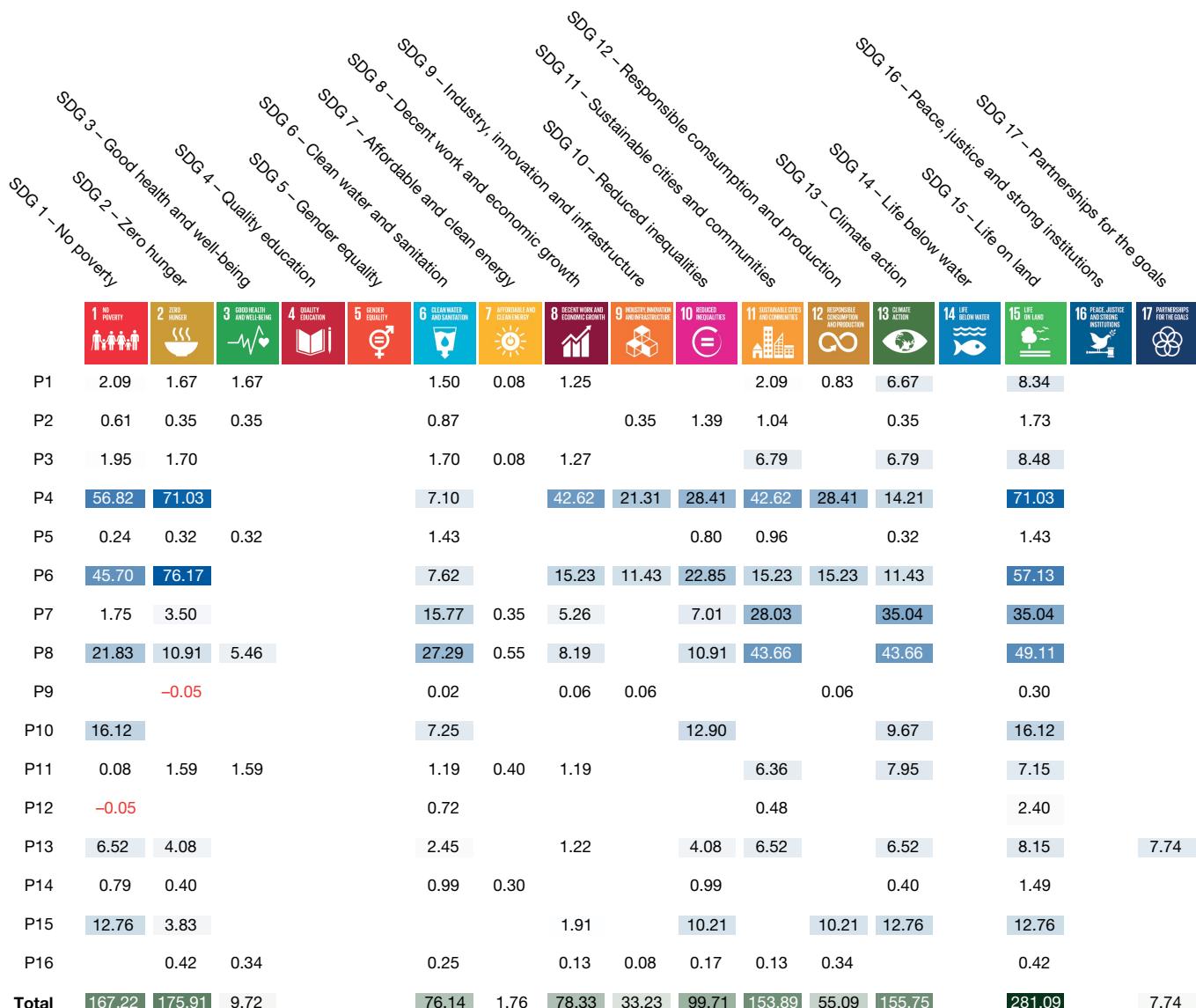


Fig. 4 | Mapping investment against the SDGs. Illustrative mapping of China's sustainability investment against the 17 UN SDGs. The depth of coloured shading reflects the level of investment for individual programmes (blue) and total (green). Numbers are in billions of US\$ (2015)

the context of these global goals. Investment and actions under many programmes simultaneously aligned with multiple SDGs, contributing to 13 of the 17 goals (Fig. 4). Most investment contributed to SDG 15 (Life on land), SDG 2 (Zero hunger), and SDG 1 (No poverty). Owing to its integrated nature, China's sustainability portfolio was characterized by synergies and co-benefits across multiple SDGs, with trade-offs present but forming only a minor component (Supplementary Table 18). For example, Comprehensive Agricultural Development Program investment directly addressed SDGs 15, 2, and 1, but also had co-benefits for several other SDGs including SDG 8 (Decent work and economic growth). Illustrating trade-offs, while Grain-for-Green impacts on water quality made a strong positive contribution to SDG 6 (Clean water and sanitation), water use by inappropriate afforestation in drylands made a minor negative contribution.

Some keys to success

The impacts of China's sustainability programmes have been heterogeneous and nuanced but overwhelmingly positive across multiple aspects of land-system sustainability. While the challenges in programme implementation have been well documented^{8,10} (Supplementary

US\$). As investment may be aligned with multiple SDGs, row totals are not presented. See Supplementary Methods for a detailed rationale behind programme investment allocation. SDG icon images courtesy of the United Nations.

Discussion), we propose some key characteristics of success and discuss the implications for other nations pursuing sustainability³.

Long-term investment at scale

Long-term investment at scale
Steadfast government commitment to large financial investment, sustained over decades, was a prerequisite for implementing the scale of interventions required to improve land-system sustainability. Massive public participation and labour was mobilized by this investment, reinforced by the establishment of conservation as a social norm¹¹¹. China's sustainability emergency demanded large-scale and long-term budgetary prioritization of addressing poverty and environmental degradation. China's experience demonstrates that achieving sustainability goals will require a step-change increase in spending for most governments¹¹², bringing it more in line with expenditure on other public services like health, defence and education. The need for long-term planning will challenge democratic governments, which typically plan over much shorter electoral cycles. For many countries, public engagement and the availability and cost of labour will also be an obstacle to improving land-system sustainability, although technology (for example, autonomous heavy machinery) may provide a partial substitute.

Systemic causes addressed

A key feature of China's programmes was that in jointly addressing systemic socio-economic and environmental causes¹¹³, the programmes aimed to break the vicious cycle of poverty and environmental degradation³⁷. Establishment of new land uses that provided both environmental and economic benefits to households underpinned the success of many programmes^{21,114}. The widespread use of financial incentives and off-farm income diversification decoupled household income from land use and reduced the need to further exploit land¹¹⁵. Parallel processes of population control, industrialization and urbanization also reduced the direct demand on China's rural land systems to provide livelihoods. In designing effective sustainability programmes, other nations must understand the complex dynamics influencing sustainability^{37,116,117} and directly target key system components, relationships and leverage points, in particular, the maintenance and diversification of landholder incomes.

A diverse, integrated portfolio

China's sustainability programmes addressed multiple sustainability challenges via a diverse range of policy instruments. Unintended consequences associated with large-scale restoration and bans on deforestation and grazing included reduced food production and timber supply just when demand was booming (Fig. 1). However, these tensions were anticipated and skilfully managed via complementary programmes aimed at increasing agricultural and timber production (although the impacts of China's timber demand are now largely outsourced to other countries via trade¹²). Programmes combined multiple policy instruments (incentives, regulation and education) and were complemented by other institutional arrangements such as population control, environmental laws, and property rights reform (Fig. 1). Despite this integrated approach, trade-offs for water⁸⁰ and biodiversity⁸⁶ were still prevalent. Sustainability portfolios must be diverse and integrated, completely addressing all aspects of sustainability to anticipate and manage trade-offs and avoid unintended consequences^{113,116–118}, including indirect impacts on other nations¹⁵.

Evidence-based, coordinated, adaptive

The response to China's sustainability emergency was required before sustainability science and landscape-scale adaptive management were well developed¹¹⁹. Nonetheless, interventions were evidence-based, informed by extensive concurrent experimentation, research and technological development^{48–50}. New information and technology were disseminated to landholders via large-scale extension and education/training⁹⁷. Programme governance has been adaptive¹²⁰, although maladaptation has occurred (for example, afforestation with inappropriate species). Pilots/trials and staged rollouts were employed to enhance programme success and, supported by local, provincial and central government coordination, ongoing project monitoring and evaluation informed regular planning and revision (Supplementary Tables 1–16, Fig. 3). Priorities were changed as capacity, knowledge and technology developed, mistakes were learned from, and sustainability objectives were achieved. Adaptive governance—where programme objectives, policy instruments and on-ground methods are highly responsive to change¹¹³—supported by multiscale coordination, is essential for the success of sustainability programmes in land-systems where complexity and uncertainty are the norm^{120,121}.

Decisive action

Deep poverty and poor living standards often co-existed with dire environmental degradation and/or risk from natural hazards. Regulatory controls in combination with ecological migration were effectively used as an emergency response to swiftly break this vicious cycle, yet the latter often caused socio-economic and cultural upheaval^{45,107,109}. Relocation continues to be part of the sustainability solution in the poorest and most environmentally degraded areas⁴⁶, although its use as a sustainability intervention seems to be declining in China¹²². Environmental regulation is already widely used to promote land-system sustainability in most nations and will remain an important tool for improving

land-system sustainability. However, if ecological migration is to be considered as a sustainability emergency response, adequate socio-economic and cultural support is essential. Approaches such as local governance and co-management, where local people are genuine partners in sustainability, may provide less culturally disruptive alternatives¹²³.

Future risks and their management

Material risks are posed to the durability of China's sustainability interventions. When payments cease, sustainability interventions are at risk of reconversion, and degradation may resume as households return to farming to supplement incomes^{79,114,124}. Forest loss from land reconversion has already been detected by remote sensing¹²⁵. Desertification interventions are at risk too¹²⁶, as evidenced by payment reductions driving herders to graze restricted grasslands and restock herds¹²⁷. In addition, urban and agricultural land uses, fuelled by increasing population and consumption¹³, are competing for land with restored forests and grasslands, and accelerating environmental change will also affect these large-scale sustainability interventions¹²⁸. Without doubt, new risks unforeseen here will also emerge. From on past experience in China's sustainability programmes, we suggest some strategies for managing these risks in uncertain situations.

Income maintenance and diversification

Maintenance and diversification of household incomes are key to minimizing the risk of reconversion of sustainability interventions¹²⁹. Many sustainability interventions were intended to generate income over time, thereby decreasing the likelihood of reconversion once initial payments end¹³⁰. Policies supporting the transition of rural labour to off-farm work can also lower reconversion rates via lessening household reliance on farm profitability and reducing labour surpluses¹²⁴. Households affected by sustainability interventions which reduce economic returns from land (such as grazing bans and forest management) that have not successfully transitioned to off-farm work¹⁰⁶ may require long-term income support or greater assistance in finding new employment. Indefinite government payments are one option, potentially justifiable by the substantial market (for example, carbon value¹³¹) and/or non-market value¹³² of the ecosystem services provided by sustainability interventions. The private sector is another potential source of long-term funding for landholder payments as Chinese citizens increasingly demand greater corporate social responsibility¹⁴.

Better planning and targeting

Long-term durability of sustainability interventions can be enhanced by better planning and targeting to capture win-win socio-economic and environmental opportunities or at least minimize trade-offs¹¹⁸. For example, the targeting of steep land, which is also less productive and less profitable, by the Grain for Green Program, minimized the opportunity cost for farmers¹³³ and reduces the likelihood of reconversion¹²⁹. Planning and targeting should also consider environmental conditions and land-use demand both current and future, identifying appropriate restoration sites, species, and ecosystems that are robust to future environmental change⁴⁸. China's thirteenth 5-year plan¹²² introduced a spatial zoning approach where specific areas are designated for agricultural production, forest conservation and development, urban development, and key ecological functions¹³⁴. This can help manage pressure from land-use competition at a large scale, especially when complemented by local-level planning.

Institutional reform and local engagement

Institutional reform and local engagement are also important for ensuring that China's sustainability interventions endure. Better consideration of local heterogeneity and dynamics in social, economic, and environmental systems in policy formulation and implementation is fundamental to widespread success and durability^{121,135}. Households with greater autonomy have reported lower reconversion intentions after programme payments cease¹²⁹. Better local engagement requires measures such as locally adapted science and technology transfer to

households¹³⁶, the tailoring of programmes to specific needs^{46,108} and effective and visible local monitoring¹²⁴. Local institutions are needed that implement national and regional sustainability priorities at the local level, and provide feedback to local, regional and national governments, in an ongoing process of multi-scale adaptive improvement¹³⁷. Stronger regulation of land subjected to state-funded sustainability interventions, and enhanced land tenure security, empowered by legal certification and stronger contractual rights, are also important for ensuring long-lasting sustainability outcomes¹³⁸.

Outlook for China and the world

China's pursuit of land-system sustainability represents a remarkable achievement of governance, policy and human endeavour. Current evidence suggests that, despite some adverse outcomes, China's integrated portfolio of sustainability programmes has achieved considerable overall success, with measurable benefits for sustainable wellbeing. However, the specific impacts of sustainability programmes are often clouded by the confounding effects of multiple socio-economic, policy and environmental factors operating concurrently^{9,59}. For example, in addition to China's sustainability programmes, economic development, industrialization and urbanization have also played important parts in improving farm-household income and reducing the pressure on land to provide livelihoods. Now, 20 years on from China's great acceleration in sustainability investment, a comprehensive and robust quantitative evaluation of the impacts of its programmes is needed. Evaluations of the specific sustainability outcomes of individual programmes are required at multiple scales using causal methods to develop reliable counterfactuals, control for confounding factors, quantify additionality, and attribute causality⁶⁶. Systematic reviews and meta-analyses are then required to quantify the nationwide impacts of individual programmes on specific indicators (for example, ref. ¹³⁹), which then need to be synthesized in a comprehensive evaluation of each programme (for example, see refs ^{5,102}). Disentangling and understanding the contribution of China's programmes to land-system sustainability is critical for informing effective adaptation of priorities and approaches. China still faces enormous social and environmental sustainability challenges associated with rapid development, industrialization and urbanization. Lessons from its land-system portfolio can also guide its approach to these broader sustainability issues such as urban pollution¹⁴ and coastal reclamation¹⁴⁰.

Similarly, China's experience in rural land systems can help other nations to contribute to humanity's shared global sustainability aspirations as embodied in the UN SDGs. Insights from China's programmes suggest that, to have a meaningful impact, other nations must start to think about sustainability as a long-term, large-scale public investment comparable to other services, such as education and infrastructure. Interventions must embrace the complex nature of sustainability challenges and target key systemic causes and leverage points, in particular, addressing socio-economic and environmental feedbacks. An integrated portfolio is required that addresses all components of land-system sustainability, anticipates and manages the trade-offs prevalent in land-systems, and avoids unintended consequences. Programmes must be evidence-based, prioritizing cost-effective interventions, with priorities and approaches adapting over time. Not only do nations need to be prepared to take urgent, strong and decisive action where required to ensure the sustainability of the environment and the people that depend on it, but also to provide appropriate socio-economic and cultural support for those affected. Given that each sustainability programme will probably contribute towards several SDGs and each SDG will be addressed by several programmes, portfolio-wide planning is essential to efficiently and effectively achieve multiple SDGs.

Received: 27 June 2017; Accepted: 16 May 2018;

Published online 11 July 2018.

1. Marks, R. B. *China: An Environmental History* (ed. Lanham, M. A.) 2nd edn (Rowman and Littlefield, Lanham, 2017). **Key, authoritative, and recently updated account of China's environmental history.**

2. Xu, J., Yin, R., Li, Z. & Liu, C. China's ecological rehabilitation: unprecedented efforts, dramatic impacts, and requisite policies. *Ecol. Econ.* **57**, 595–607 (2006). **Early overview and background of the Grain for Green Program and the Natural Forest Conservation Program and assessment of their challenges.**
3. United Nations. *Transforming Our World: The 2030 Agenda for Sustainable Development Annex A/RES/70/1. https://sustainabledevelopment.un.org/post2015/transformingourworld* (UN, 2015).
4. Gao, L. & Bryan, B. A. Finding pathways to national-scale land-sector sustainability. *Nature* **544**, 217–222 (2017).
5. Delang, C. O. & Yuan, Z. *China's Grain for Green Program: A Review of the Largest Ecological Restoration and Rural Development Program in the World* (Springer International Publishing, Switzerland, 2015). **Deep review of the Grain for Green Program which thoroughly covers multiple environmental, policy, and socio-economic details.**
6. Yin, R. *An Integrated Assessment of China's Ecological Restoration Programs* (Springer, East Lansing, 2009). **Collection of articles describing several of China's sustainability programmes.**
7. Liu, J., Ouyang, Z., Yang, W., Xu, W. & Li, S. in *Encyclopedia of Biodiversity* Vol. 3, 372–384 (Academic Press, Waltham, 2013).
8. Liu, J., Li, S., Ouyang, Z., Tam, C. & Chen, X. Ecological and socioeconomic effects of China's policies for ecosystem services. *Proc. Natl. Acad. Sci. USA* **105**, 9477–9482 (2008). **High-profile critical assessment that raised awareness of the scale of China's investment in sustainability.**
9. Ouyang, Z. Y. et al. Improvements in ecosystem services from investments in natural capital. *Science* **352**, 1455–1459 (2016). **National-scale quantitative assessment of the changes in ecosystem services across China, relating these to sustainability investment.**
10. Yin, R. S., Yin, G. P. & Li, L. Y. Assessing China's ecological restoration programs: what's been done and what remains to be done? *Environ. Manage.* **45**, 442–453 (2010).
11. Yin, R. S. & Yin, G. P. China's primary programs of terrestrial ecosystem restoration: initiation, implementation, and challenges. *Environ. Manage.* **45**, 429–441 (2010).
12. Liu, J. G. & Diamond, J. China's environment in a globalizing world. *Nature* **435**, 1179–1186 (2005).
13. Liu, J. & Raven, P. H. China's environmental challenges and implications for the world. *Crit. Rev. Environ. Sci. Technol.* **40**, 823–851 (2010).
14. Kahn, M. E. & Zheng, S. *Blue Skies over Beijing: Economic Growth and the Environment in China* (Princeton Univ. Press, Princeton, 2016).
15. Shapiro, J. *China's Environmental Challenges* (Wiley, Polity Press, Cambridge, 2016).
16. Banister, J. in *The Population of Modern China* (eds Poston, D. L. & D. Yaukey, D.) 51–57 (Springer, Boston, 1992).
17. He, F., Ge, Q., Dai, J. & Rao, Y. Forest change of China in recent 300 years. *J. Geogr. Sci.* **18**, 59–72 (2008).
18. Elvin, M. *The Retreat of the Elephants: An Environmental History of China* (Yale Univ. Press, New Haven, 2004).
19. Mao, Y., Zhao, N. & Yang, X. in *Food Security and Farm Land Protection in China* Vol. 2 (eds Yang, M. & Fan, G.) 356 (Series on Chinese Economics Research, World Scientific Publishing, Singapore, 2013).
20. Miao, L. et al. Synthesis of China's land use in the past 300 years. *Global Planet. Change* **100**, 224–233 (2013). **Comprehensive synthesis of land-use and population dynamics in China, combining multiple datasets.**
21. Miao, L., Zhu, F., Sun, Z., Moore, J. & Cui, X. China's land-use changes during the past 300 years: a historical perspective. *Int. J. Environ. Res. Public Health* **13**, 847 (2016).
22. Hua, L. M. & Squires, V. R. Managing China's pastoral lands: current problems and future prospects. *Land Use Policy* **43**, 129–137 (2015).
23. Liu, J. & Yang, W. Water sustainability for China and beyond. *Science* **337**, 649–650 (2012).
24. Yu, D. P. et al. Forest management in northeast China: history, problems, and challenges. *Environ. Manage.* **48**, 1122–1135 (2011).
25. Ren, G. et al. Effectiveness of China's National Forest Protection Program and nature reserves. *Conserv. Biol.* **29**, 1368–1377 (2015).
26. Xu, J. China's new forests aren't as green as they seem. *Nature* **477**, 371 (2011). **Opinion piece challenging the environmental credentials of China's large-scale reforestation and afforestation programmes.**
27. Ran, L. S., Lu, X. X. & Xu, J. C. Effects of vegetation restoration on soil conservation and sediment loads in China: a critical review. *Crit. Rev. Environ. Sci. Technol.* **43**, 1384–1415 (2013).
28. Lei, J. & Zhu, L. China's Implementation of Six Key Forestry Programs. <http://www.china.org.cn/e-news/news02-05-14.htm> (China Tibet Information Center/State Forestry Administration, 2002).
29. Douglas, I. Land degradation, soil conservation and the sediment load of the Yellow River, China: review and assessment. *Land Degrad. Rehabil.* **1**, 141–151 (1989).
30. Li, X. Y. et al. Patterns of cereal yield growth across China from 1980 to 2010 and their implications for food production and food security. *PLoS One* **11**, e0159061 (2016).
31. PRC Ministry of Agriculture. *Notification on National Farmland Quality Grading* (Beijing, China, 2014) [in Chinese].
32. Chen, Y. & Tang, H. Desertification in north China: background, anthropogenic impacts and failures in combating it. *Land Degrad. Dev.* **16**, 367–376 (2005).

33. Wang, F., Pan, X., Wang, D., Shen, C. & Lu, Q. Combating desertification in China: Past, present and future. *Land Use Policy* **31**, 311–313 (2013).

34. Feng, Q., Ma, H., Jiang, X. M., Wang, X. & Cao, S. X. What has caused desertification in China? *Sci. Rep.* **5**, 15998 (2015).

35. Wang, X., Chen, F., Hasi, E. & Li, J. Desertification in China: an assessment. *Earth Sci. Rev.* **88**, 188–206 (2008).

36. Xie, X. H. et al. Detection and attribution of changes in hydrological cycle over the Three-North region of China: climate change versus afforestation effect. *Agric. For. Meteorol.* **203**, 74–87 (2015).

37. Cao, S. X., Zhong, B. L., Yue, H., Zeng, H. S. & Zeng, J. H. Development and testing of a sustainable environmental restoration policy on eradicating the poverty trap in China's Changting County. *Proc. Natl. Acad. Sci. USA* **106**, 10712–10716 (2009).

Identifies the importance of a systemic approach and the joint solution of poverty and environmental degradation.

38. Cheng, L. et al. Estimation of the costs of desertification in China: a critical review. *Land Degrad. Dev.* **29**, 975–983 (2016).

39. Shiau, J.-T., Feng, S. & Nadarajah, S. Assessment of hydrological droughts for the Yellow River, China, using copulas. *Hydrol. Processes* **21**, 2157–2163 (2007).

40. Ye, Q. & Glantz, M. H. The 1998 Yangtze Floods: the use of short-term forecasts in the context of seasonal to interannual water resource management. *Mitigation Adapt. Strategies Glob. Change* **10**, 159–182 (2005).

41. Wang, X., Dong, Z., Zhang, J. & Liu, L. Modern dust storms in China: an overview. *J. Arid Environ.* **58**, 559–574 (2004).

42. Ai, N. & Polenski, K. R. Socioeconomic impact analysis of yellow-dust storms: an approach and case study for Beijing. *Econ. Syst. Res.* **20**, 187–203 (2008).

43. Farm Service Agency. *Conservation Reserve Program Statistics*. <https://www.fsa.usda.gov/programs-and-services/conservation-programs/reports-and-statistics/conservation-reserve-program-statistics/index> (United States Department of Agriculture, FSA, 2017).

44. Yang, W. et al. Performance and prospects of payments for ecosystem services programs: evidence from China. *J. Environ. Manage.* **127**, 86–95 (2013).

45. Dong, C., Liu, X. M. & Klein, K. K. Land degradation and population relocation in northern China. *Asia Pacif. Viewp.* **53**, 163–177 (2012).

46. Wang, P. et al. Promise and reality of market-based environmental policy in China: empirical analyses of the ecological restoration program on the Qinghai-Tibetan Plateau. *Glob. Environ. Change* **39**, 35–44 (2016).

47. Lei, Y. R., Finlayson, C. M., Thwaites, R., Shi, G. Q. & Cui, L. J. Using government resettlement projects as a sustainable adaptation strategy for climate change. *Sustainability* **9**, 1373 (2017).

48. Ghestem, M. et al. A framework for identifying plant species to be used as 'ecological engineers' for fixing soil on unstable slopes. *PLoS One* **9**, e95876 (2014).

49. Li, X. R., Xiao, H. L., He, M. Z. & Zhang, J. G. Sand barriers of straw checkerboards for habitat restoration in extremely arid desert regions. *Ecol. Eng.* **28**, 149–157 (2006).

50. Yang, Z.-Q., Wang, X.-Y. & Zhang, Y.-N. Recent advances in biological control of important native and invasive forest pests in China. *Biol. Control* **68**, 117–128 (2014).

51. Li, M.-M. et al. An overview of the "Three-North" Shelterbelt project in China. *For. Stud. China* **14**, 70–79 (2012).

52. Liao, C., Han, F. & Feng, M. Construction achievement and experience of soil and water conservation. *Yangtze River* **41**, 16–20 (2010) [in Chinese].

53. PRC Information Office of the State Council. *New Progress in Development-oriented Poverty Reduction Program for Rural China*. http://www.gov.cn/english/official/2011-11/16/content_1994729.htm (PRCIO, Beijing, 2011).

54. Li, P. & Wang, X. in *Ecological Migration, Development and Transformation: A Study of Migration and Poverty Reduction in Ningxia* (eds Li, P. & Wang, X.) 1–19 (Springer, Berlin, 2016).

55. Wang, J. G. Review of the Comprehensive Agricultural Development Program after two decades of development. *China State Finance* **18**, 32–34 (2008) [in Chinese].

56. The World Bank. *Forest area (% of land area): China*. <https://data.worldbank.org/indicator/AG.LND.FRST.ZS?locations=CN> (The World Bank, 2017).

57. Chen, Y. et al. Balancing green and grain trade. *Nat. Geosci.* **8**, 739–741 (2015).

58. Yang, H. F., Mu, S. J. & Li, J. L. Effects of ecological restoration projects on land use and land cover change and its influences on territorial NPP in Xinjiang, China. *Catena* **115**, 85–95 (2014).

59. Lu, Y. H. et al. Recent ecological transitions in China: greening, browning, and influential factors. *Sci. Rep.* **5**, 8732 (2015).

Quantifies the complex spatial distribution of greening/browning across China from 2000–2010 and the influence of sustainability interventions.

60. Viña, A., McConnell, W. J., Yang, H., Xu, Z. & Liu, J. Effects of conservation policy on China's forest recovery. *Sci. Adv.* **2**, e1500965 (2016).

61. Huang, L. et al. Effects of grassland restoration programs on ecosystems in arid and semiarid China. *J. Environ. Manage.* **117**, 268–275 (2013).

62. Xiong, D. P., Shi, P. L., Zhang, X. Z. & Zou, C. B. Effects of grazing exclusion on carbon sequestration and plant diversity in grasslands of China: a meta-analysis. *Ecol. Eng.* **94**, 647–655 (2016).

63. Mu, S. J. et al. Assessing the impact of restoration-induced land conversion and management alternatives on net primary productivity in Inner Mongolian grassland, China. *Global Planet. Change* **108**, 29–41 (2013).

64. Cai, H. Y., Yang, X. H. & Xu, X. L. Human-induced grassland degradation/restoration in the central Tibetan Plateau: the effects of ecological protection and restoration projects. *Ecol. Eng.* **83**, 112–119 (2015).

65. Piao, S., Fang, J., Liu, H. & Zhu, B. NDVI-indicated decline in desertification in China in the past two decades. *Geophys. Res. Lett.* **32**, L06402 (2005).

66. Wang, X. M., Zhang, C. X., Hasi, E. & Dong, Z. B. Has the Three Norths Forest Shelterbelt Program solved the desertification and dust storm problems in arid and semiarid China? *J. Arid Environ.* **74**, 13–22 (2010).

Critical review finding little unsatisfactory evidence supporting the impact of afforestation programmes on desertification and dust storm mitigation in China and calls for stronger causal analyses.

67. Peng, D. L. et al. The influences of drought and land-cover conversion on inter-annual variation of NPP in the Three-North Shelterbelt Program zone of China based on MODIS data. *PLoS One* **11**, e0158173 (2016).

68. Zhang, Y. et al. Multiple afforestation programs accelerate the greenness in the 'Three North' region of China from 1982 to 2013. *Ecol. Indic.* **61**, 404–412 (2016).

69. Li, X. S., Wang, H. Y., Wang, J. Y. & Gao, Z. H. Land degradation dynamic in the first decade of twenty-first century in the Beijing-Tianjin dust and sandstorm source region. *Environ. Earth Sci.* **74**, 4317–4325 (2015).

70. Deng, L., Shangguan, Z.-P. & Li, R. Effects of the Grain-for-Green program on soil erosion in China. *Int. J. Sediment Res.* **27**, 120–127 (2012).

71. Li, C. B. et al. Quantifying the effect of ecological restoration on soil erosion in China's Loess Plateau region: an application of the MMF approach. *Environ. Manage.* **45**, 476–487 (2010).

72. Zhang, J., Wang, T. & Ge, J. Assessing vegetation cover dynamics induced by policy-driven ecological restoration and implication to soil erosion in southern China. *PLoS One* **10**, e0131352 (2015).

73. Liu, Y. S., Guo, Y. J., Li, Y. R. & Li, Y. H. GIS-based effect assessment of soil erosion before and after gully land consolidation: a case study of Wangjiagou project region, Loess Plateau. *Chin. Geogr. Sci.* **25**, 137–146 (2015).

74. Wang, S. A. et al. Reduced sediment transport in the Yellow River due to anthropogenic changes. *Nat. Geosci.* **9**, 38–41 (2016).

75. Yang, S. L., Xu, K. H., Milliman, J. D., Yang, H. F. & Wu, C. S. Decline of Yangtze River water and sediment discharge: impact from natural and anthropogenic changes. *Sci. Rep.* **5**, 12581 (2015).

76. An, W. M. et al. Exploring the effects of the "Grain for Green" program on the differences in soil water in the semi-arid Loess Plateau of China. *Ecol. Eng.* **107**, 144–151 (2017).

77. Cao, S. X. et al. Excessive reliance on afforestation in China's arid and semi-arid regions: lessons in ecological restoration. *Earth Sci. Rev.* **104**, 240–245 (2011).

78. Lu, C., Zhao, T., Shi, X. & Cao, S. Ecological restoration by afforestation may increase groundwater depth and create potentially large ecological and water opportunity costs in arid and semiarid China. *J. Clean. Prod.* **176**, 1213–1222 (2018).

79. Cao, S. Impact of China's large-scale ecological restoration program on the environment and society in arid and semiarid areas of China: achievements, problems, synthesis, and applications. *Crit. Rev. Environ. Sci. Technol.* **41**, 317–335 (2011).

80. Cao, S. X., Zhang, J. Z., Chen, L. & Zhao, T. Y. Ecosystem water imbalances created during ecological restoration by afforestation in China, and lessons for other developing countries. *J. Environ. Manage.* **183**, 843–849 (2016).

81. Feng, X. et al. Revegetation in China's Loess Plateau is approaching sustainable water resource limits. *Nat. Clim. Chang.* **6**, 1019–1022 (2016).

Quantification of the impacts of reforestation and afforestation and the critical state of water resources in the Loess Plateau region.

82. Xu, W. et al. Strengthening protected areas for biodiversity and ecosystem services in China. *Proc. Natl. Acad. Sci. USA* **114**, 1601–1606 (2017).

Recent assessment of the representativeness of China's nature reserve system and opportunities for improvement identifying the importance of payment schemes in shifting rural labour off-farm and reducing direct reliance on natural resources for livelihoods.

83. Zhang, K. R. et al. Sustainability of social-ecological systems under conservation projects: lessons from a biodiversity hotspot in western China. *Biol. Conserv.* **158**, 205–213 (2013).

84. Zhai, D. L., Xu, J. C., Dai, Z. C., Cannon, C. H. & Grumbine, R. E. Increasing tree cover while losing diverse natural forests in tropical Hainan, China. *Reg. Environ. Change* **14**, 611–621 (2014).

85. Wang, X. L., Wang, Y. Q. & Wang, Y. J. Use of exotic species during ecological restoration can produce effects that resemble vegetation invasions and other unintended consequences. *Ecol. Eng.* **52**, 247–251 (2013).

86. Hua, F. et al. Opportunities for biodiversity gains under the world's largest reforestation programme. *Nat. Commun.* **7**, 12717 (2016).

87. Wang, S. L., Tuan, F., Gale, F., Somwaru, A. & Hansen, J. China's regional agricultural productivity growth in 1985–2007: a multilateral comparison. *Agric. Econ.* **44**, 241–251 (2013).

88. Yang, H. Livestock development in China: animal production, consumption and genetic resources. *J. Anim. Breed. Genet.* **130**, 249–251 (2013).

89. Chen, Y.-f. et al. Agricultural policy, climate factors and grain output: evidence from household survey data in rural China. *J. Integr. Agric.* **12**, 169–183 (2013).

90. Zhang, J. China's success in increasing per capita food production. *J. Exp. Bot.* **62**, 3707–3711 (2011).

91. Lassaleta, L., Billen, G., Grizzetti, B., Anglade, J. & Garnier, J. 50 year trends in nitrogen use efficiency of world cropping systems: the relationship between yield and nitrogen input to cropland. *Environ. Res. Lett.* **9**, 105011 (2014).

92. Maryna, S. et al. Alarming nutrient pollution of Chinese rivers as a result of agricultural transitions. *Environ. Res. Lett.* **11**, 024014 (2016).

93. Le, C. et al. Eutrophication of lake waters in China: cost, causes, and control. *Environ. Manage.* **45**, 662–668 (2010).

94. Jin, X. et al. The evaluation of land consolidation policy in improving agricultural productivity in China. *Sci. Rep.* **7**, 2792 (2017).

95. Song, W. & Pijanowski, B. C. The effects of China's cultivated land balance program on potential land productivity at a national scale. *Appl. Geogr.* **46**, 158–170 (2014).

96. Yan, H., Liu, J., Huang, H. Q., Tao, B. & Cao, M. Assessing the consequence of land use change on agricultural productivity in China. *Global Planet. Change* **67**, 13–19 (2009).

97. Yao, S. & Li, H. Agricultural productivity changes induced by the Sloping Land Conversion Program: an analysis of Wuqi County in the Loess Plateau region. *Environ. Manage.* **45**, 541–550 (2010).

98. Lu, Q., Xu, B., Liang, F., Gao, Z. & Ning, J. Influences of the Grain-for-Green project on grain security in southern China. *Ecol. Indic.* **34**, 616–622 (2013).

99. Yin, R. S., Liu, C., Zhao, M. J., Yao, S. B. & Liu, H. The implementation and impacts of China's largest payment for ecosystem services program as revealed by longitudinal household data. *Land Use Policy* **40**, 45–55 (2014).

100. Yi, F. J., Sun, D. Q. & Zhou, Y. H. Grain subsidy, liquidity constraints and food security—impact of the grain subsidy program on the grain-sown areas in China. *Food Policy* **50**, 114–124 (2015).

101. Liu, C., Mullan, K., Liu, H., Zhu, W. Q. & Rong, Q. J. The estimation of long term impacts of China's key priority forestry programs on rural household incomes. *J. For. Econ.* **20**, 267–285 (2014).

102. Gutiérrez Rodríguez, L. et al. China's conversion of cropland to forest program: a systematic review of the environmental and socioeconomic effects. *Environ. Evid.* **5**, 21 (2016).

103. Liu, C., Lu, J. Z. & Yin, R. S. An estimation of the effects of China's priority forestry programs on farmers' income. *Environ. Manage.* **45**, 526–540 (2010).

104. Uchida, E., Rozelle, S. & Xu, J. Conservation payments, liquidity constraints, and off-farm labor: impact of the Grain-for-Green Program on rural households in China. *Am. J. Agric. Econ.* **91**, 70–86 (2009).

105. Li, H., Yao, S. B., Yin, R. S. & Liu, G. Q. Assessing the decadal impact of China's Sloping Land Conversion Program on household income under enrollment and earning differentiation. *For. Policy Econ.* **61**, 95–103 (2015).

106. Cao, S., Wang, X., Song, Y., Chen, L. & Feng, Q. Impacts of the Natural Forest Conservation Program on the livelihoods of residents of northwestern China: perceptions of residents affected by the program. *Ecol. Econ.* **69**, 1454–1462 (2010).

107. Wang, Z., Song, K. & Hu, L. China's largest scale ecological migration in the Three-River Headwater region. *Ambio* **39**, 443–446 (2010).

108. Shu, X. in *Ecological Migration, Development and Transformation: A Study of Migration and Poverty Reduction in Ningxia* (eds Li, P. & Wang, X.) 21–46 (Springer, Berlin, 2016).

109. Xie, Y. *Ecological Migrants: The Relocation of China's Ewenki Reindeer Herders* (Berghahn Books, New York, 2015).

110. Mao, X. F., Wei, X. Y. & Xia, J. X. Evaluation of ecological migrants' adaptation to their new living area in Three-River Headwater wetlands, China. *Proc. Environ. Sci.* **13**, 1346–1353 (2012).

111. Chen, X. D., Lupi, F., He, G. M. & Liu, J. G. Linking social norms to efficient conservation investment in payments for ecosystem services. *Proc. Natl Acad. Sci. USA* **106**, 11812–11817 (2009).

112. UNCTAD. *World Investment Report 2014. Investing in the SDGs: An Action Plan*. http://unctad.org/en/PublicationsLibrary/wir2014_en.pdf (United Nations Conference on Trade and Development, Switzerland, 2014).

113. Liu, J. G. et al. Complexity of coupled human and natural systems. *Science* **317**, 1513–1516 (2007).

114. Cao, S. X., Shang, D., Yue, H. & Ma, H. A win-win strategy for ecological restoration and biodiversity conservation in southern China. *Environ. Res. Lett.* **12**, 044004 (2017).

115. Li, T. et al. Gauging policy-driven large-scale vegetation restoration programmes under a changing environment: their effectiveness and socio-economic relationships. *Sci. Total Environ.* **607**, 911–919 (2017).

116. Liu, J. et al. *Pandas and People: Coupling Human and Natural Systems for Sustainability* (Oxford Univ. Press, Oxford, 2016).

117. Liu, J. et al. Systems integration for global sustainability. *Science* **347**, 1258832 (2015).

118. Bryan, B. A. et al. Land use efficiency: anticipating future demand for land-sector greenhouse gas emissions abatement and managing trade-offs with agriculture, water, and biodiversity. *Glob. Change Biol.* **21**, 4098–4114 (2015).

119. Kates, R. W. et al. Sustainability science. *Science* **292**, 641–642 (2001).

120. Schultz, L., Folke, C., Österblom, H. & Olsson, P. Adaptive governance, ecosystem management, and natural capital. *Proc. Natl Acad. Sci. USA* **112**, 7369–7374 (2015).

121. Dietz, T., Ostrom, E. & Stern, P. C. The struggle to govern the commons. *Science* **302**, 1907–1912 (2003).

122. Central Committee of the Communist Party of China. *The 13th Five-Year Plan For Economic and Social Development of the People's Republic of China 2016–2020*. <http://en.ndrc.gov.cn/newsrelease/201612/P020161207645765233498.pdf> (CCP, Beijing, 2015).

123. Foggin, J. M. Rethinking "ecological migration" and the value of cultural continuity: a response to Wang, Song, and Hu. *Ambio* **40**, 100–101 (2011).

124. Song, C. H. et al. Sustainability of forests created by China's Sloping Land Conversion Program: a comparison among three sites in Anhui, Hubei and Shanxi. *For. Policy Econ.* **38**, 161–167 (2014).

125. Guo, J. & Gong, P. Forest cover dynamics from Landsat time-series data over Yan'an city on the Loess Plateau during the Grain for Green Project. *Int. J. Remote Sens.* **37**, 4101–4118 (2016).

126. Liu, N., Zhou, L. H. & Hauger, J. S. How sustainable is government-sponsored desertification rehabilitation in China? Behavior of households to changes in environmental policies. *PLoS One* **8**, e77510 (2013).

127. Zhen, L. et al. Herders' willingness to accept versus the public sector's willingness to pay for grassland restoration in the Xilingol League of Inner Mongolia, China. *Environ. Res. Lett.* **9**, 045003 (2014).

128. He, B., Chen, A. F., Wang, H. L. & Wang, Q. F. Dynamic response of satellite-derived vegetation growth to climate change in the Three North Shelter Forest region in China. *Remote Sens.* **7**, 9998–10016 (2015).

129. Yang, X. J. & Xu, J. T. Program sustainability and the determinants of farmers' self-predicted post-program land use decisions: evidence from the Sloping Land Conversion Program (SLCP) in China. *Environ. Dev. Econ.* **19**, 30–47 (2014).

130. Frayer, J., Sun, Z. L., Muller, D., Munroe, D. K. & Xu, J. C. Analyzing the drivers of tree planting in Yunnan, China, with Bayesian networks. *Land Use Policy* **36**, 248–258 (2014).

131. Wang, Z. X. & Lu, Y. Compensation for the conversion of sloping farmland to forest in China: a feasibility study of payment based on carbon sink. *J. Environ. Dev.* **19**, 28–41 (2010).

132. Wang, X., Bennett, J., Xie, C., Zhang, Z. & Liang, D. Estimating non-market environmental benefits of the Conversion of Cropland to Forest and Grassland Program: a choice modeling approach. *Ecol. Econ.* **63**, 114–125 (2007).

133. Kelly, P. & Huo, X. X. Do farmers or governments make better land conservation choices? Evidence from China's Sloping Land Conversion Program. *J. For. Econ.* **19**, 32–60 (2013).

134. Lu, Y. H., Ma, Z. M., Zhang, L. W., Fu, B. J. & Gao, G. Y. Redlines for the greening of China. *Environ. Sci. Policy* **33**, 346–353 (2013).

135. Ostrom, E. A general framework for analyzing sustainability of social-ecological systems. *Science* **325**, 419–422 (2009).

136. Zhang, W. et al. Closing yield gaps in China by empowering smallholder farmers. *Nature* **537**, 671–674 (2016).

137. He, J. Governing forest restoration: local case studies of Sloping Land Conversion Program in southwest China. *For. Policy Econ.* **46**, 30–38 (2014).

138. Yi, Y. Y., Kohlin, G. & Xu, J. T. Property rights, tenure security and forest investment incentives: evidence from China's Collective Forest Tenure Reform. *Environ. Dev. Econ.* **19**, 48–73 (2014).

139. Song, X. Z., Peng, C. H., Zhou, G. M., Jiang, H. & Wang, W. F. Chinese Grain for Green Program led to highly increased soil organic carbon levels: a meta-analysis. *Sci. Rep.* **4**, 4460 (2014).

140. Ma, Z. et al. Rethinking China's new great wall. *Science* **346**, 912–914 (2014).

Acknowledgements This work was supported by a Climate Change Engagement Grant from the Australian Department of Foreign Affairs and Trade, as well as by our own institutions, in particular Deakin University. We thank M. Klaassen, D. Driscoll, J. G. Canadell and B. Huang for comments on the manuscript. This work contributes to both the Future Earth and Global Land Programme research agendas.

Reviewer information *Nature* thanks R. Costanza, F. Zhang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.A.B. designed the study and wrote the paper. L.G., Y.Y., and X.S. contributed to the writing, assembled the data and photographs, prepared the graphs, and assembled the Supplementary Information. All authors made substantive intellectual contributions to the paper and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0280-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to B.A.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Observation of half-integer thermal Hall conductance

Mitali Banerjee¹, Moty Heiblum^{1*}, Vladimir Umansky¹, Dima E. Feldman², Yuval Oreg¹ & Ady Stern¹

Topological states of matter are characterized by topological invariants, which are physical quantities whose values are quantized and do not depend on the details of the system (such as its shape, size and impurities). Of these quantities, the easiest to probe is the electrical Hall conductance, and fractional values (in units of e^2/h , where e is the electronic charge and h is the Planck constant) of this quantity attest to topologically ordered states, which carry quasiparticles with fractional charge and anyonic statistics. Another topological invariant is the thermal Hall conductance, which is harder to measure. For the quantized thermal Hall conductance, a fractional value in units of κ_0 ($\kappa_0 = \pi^2 k_B^2 / (3h)$, where k_B is the Boltzmann constant) proves that the state of matter is non-Abelian. Such non-Abelian states lead to ground-state degeneracy and perform topological unitary transformations when braided, which can be useful for topological quantum computation. Here we report measurements of the thermal Hall conductance of several quantum Hall states in the first excited Landau level and find that the thermal Hall conductance of the $5/2$ state is compatible with a half-integer value of $2.5\kappa_0$, demonstrating its non-Abelian nature.

The even-denominator fractional quantum Hall state in the first excited Landau level at a bulk filling factor of $\nu = 5/2$ has been a subject of extensive research for the past thirty years¹. After its first observation², it was suggested that this state might be a manifestation of superconducting-like condensation of composite fermions in a zero effective magnetic field^{3–5}. Furthermore, it was predicted that the state carries quasiparticles whose mutual exchange statistics is non-Abelian³. Consequently, the ground state of several quasiparticles remains degenerate even at their fixed positions; hence, such states are attractive for topological quantum computing¹. Yet, this prediction has been hard to test experimentally because relatively easily accessible experimental probes, such as electric response functions, do not reflect the topological order of the state. Even after the demonstration of the state's quasi-particle charge^{6,7} being $e^* = e/4$ and the observation of a topologically protected upstream-propagating neutral mode⁸, a family of possible orders are still viable candidates for the $\nu = 5/2$ state. However, the thermal Hall conductance may distinguish Abelian from non-Abelian states because it is quantized to an integer value of $\kappa_0 T$, where T is the temperature, for the former and a half-integer value for the latter⁵. Furthermore, its precise value distinguishes between different candidate orders. Here, we report an observation compatible with such half-integer quantization.

Background

It is worth giving a brief summary of the different orders predicted for this state. Numerical work lent support^{9–11} to the non-Abelian Pfaffian³ and anti-Pfaffian topological orders^{12,13}. Among the other possible states are the $SU(2)_2$, the $K = 8, 331$ and 113 liquids, as well as their particle-hole conjugates^{14–17}. A non-Abelian particle-hole Pfaffian order^{13,18–21} with the wave function described in ref. ¹⁹ was interpreted in terms of Dirac composite fermions¹⁸. Although the naming of these states does not follow any particular methodology, a wire construction organizes them in terms of their thermal Hall conductance, introduces possible generalizations, for which the thermal Hall conductance may be any arbitrary integer or half-integer multiple²² of $\kappa_0 T$, and identifies

their edge structure (see below). Our goal is to determine the experimentally relevant order from thermal transport.

The thermal Hall conductance is defined in a two-terminal measurement as $g_Q = dJ_Q/dT = KT$, with J_Q being the heat current (in watts) and K the thermal conductance coefficient. This highly important characteristic of the system has a maximal value for one-dimensional ballistic channels, with $K = \kappa_0$. The thermal Hall conductance is independent of the charge and the exchange statistics of the heat-carrying particles. This quantum limit has been experimentally realized for bosons^{23,24}, fermions²⁵ and recently for a strongly interacting system—the lowest Landau level in the fractional quantum Hall effect (with fractionally charged quasiparticles)²⁶. In the latter study²⁶ it was shown that in the presence of counter-propagating one-dimensional modes and in the limit of a long propagation length, the thermal conductance reflects the net number of topological chiral modes (the number of downstream modes minus the number of upstream modes), as predicted by the K matrix in the bulk²⁷.

Among the possible topological orders for the $\nu = 5/2$ state¹⁶, the non-Abelian candidates are predicted to conduct $n + 1/2$ ($n = 0, 1, \dots, 4$) units of the quantized heat, J_Q , whereas for Abelian states the $1/2$ is missing. The term $1/2$ originates from a neutral edge mode (it can be downstream or upstream) whose central charge is one-half. This mode may be viewed as a Majorana chiral edge mode of a superconductor of composite fermions. Moreover, each of the proposed non-Abelian topological orders has a different n , implying a different fractional quantized thermal conductance, g_Q . Hence, measuring the heat conductance of the equilibrated $\nu = 5/2$ fractional state may determine the nature of the topological order. Our measurement results are compatible with $K \approx 2.5\kappa_0$.

Experimental details

Our experimental setup, with its 'heart' shown in Fig. 1a, is similar in principle to our previously studied configuration²⁶ (see Methods). The two-dimensional electron gas is structured in the form of four separated arms (formed by chemical etching), with a small floating reservoir

¹Braun Center of Sub-Micron Physics, Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot, Israel. ²Department of Physics, Brown University, Providence, RI, USA.
*e-mail: moty.heiblum@weizmann.ac.il

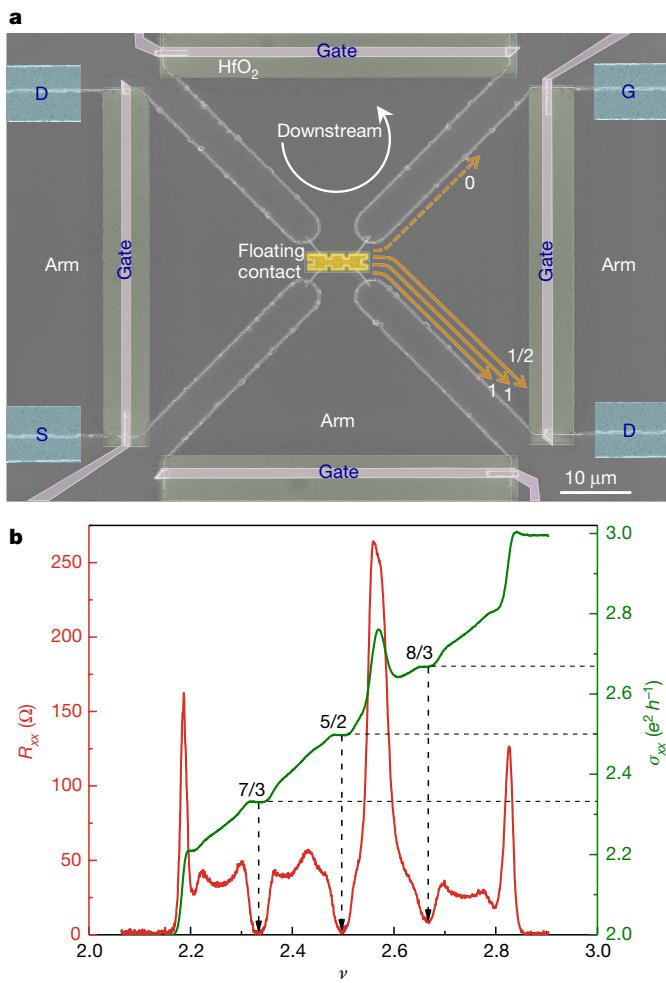


Fig. 1 | Device configuration and Hall data. **a**, The ‘heart’ of the device. For a description of the whole structure, see ref. ²⁶. The small ohmic contact (with an area of $12 \mu\text{m}^2$) serves as the heated floating reservoir, injecting currents into N arms. The effective propagation length (to a cold contact) in each arm is about $150 \mu\text{m}$. The thermal noise can be measured in two opposite arms (connected to band-pass filters and cold amplifiers). Arms can be disconnected by negatively charging surface gates deposited on HfO_2 . Although the voltage required to disconnect an arm by a typical quantum point contact is about -10 V , which leads to severe hysteresis and instability, only about -1 V sufficed with the surface gates. This allowed stable operation. As an example, the energy-carrying edge modes that correspond to the particle-hole Pfaffian order are shown. The solid orange arrows represent downstream charge modes; each carries a heat flux of $\kappa_0 T$. The dashed orange arrow represents an upstream Majorana mode carrying a heat flux of $0.5\kappa_0 T$. S, source; D, drain; G, gate. **b**, The longitudinal resistance, R_{xx} , and the transverse Hall conductance, σ_{xx} , in the first excited Landau level, measured in a separate Hall bar (length, $200 \mu\text{m}$; width, $100 \mu\text{m}$) fabricated with the same MBE-grown material as that used in the experiment.

in the centre (an ohmic contact with an area of about $12 \mu\text{m}^2$) heated by an incident direct current to temperature T_m . The chiral edge modes leave the floating reservoir and enter the separated arms with temperature T_m . Each arm is gated across its width with a continuous metallic surface gate (isolated from the sample’s surface by 5-nm-thick HfO_2), about $30 \mu\text{m}$ away from the floating reservoir. Negative charging of each gate allows disconnecting the corresponding arm from the circuit.

If I_s is the source current, then the incident current at the floating reservoir is $I_{\text{in}} = t_1 I_s$, where $t_1 = \nu_{\text{gate}}/\nu$ is the transmission coefficient of the source’s arm gate, ν_{gate} is the filling factor under the gate. The outgoing current splits into N arms, with the dissipated power in the floating reservoir being $\Delta P = P_{\text{in}} - P_{\text{out}} = 0.5 I_{\text{in}} V_s (1 - N^{-1})$, where V_s is the Hall voltage, with all gates uncharged. In thermal equilibrium,

the dissipated power equals—ideally—the outgoing power carried by the chiral one-dimensional charged edge modes and by the phonons (to the bulk); namely, $\Delta P = \Delta P_e + \Delta P_{\text{ph}}$. The edge modes are expected to carry a heat of $\Delta P_e = (1/2) n_{\text{tot}} K (T_m^2 - T_0^2)$, where $n_{\text{tot}} K$ is the overall thermal conductance coefficient of n_{tot} modes in N arms, and T_0 is the electron temperature in the grounded contacts. In turn, the heat flux carried by phonons is expected to obey²⁸ $\Delta P_{\text{ph}} = \beta (T_m^5 - T_0^5)$.

The temperature T_m was determined by measuring the thermal noise (in the downstream drain) carried by the current leaving the floating reservoir. We emphasize that owing to the chirality of the modes (and the absence of back-scattering), the low-frequency current fluctuations leaving the floating reservoir are conserved. Thus, the downstream thermal current fluctuations reflect the temperature of the floating contact (even if the edge modes cool along their paths). The voltage fluctuations in the drain, S_v (in $\text{V}^2 \text{Hz}^{-1}$) were filtered by an LC circuit located at the mixing chamber (with frequency $f_0 \approx 695 \text{ kHz}$ and bandwidth $\Delta f = 30 \text{ kHz}$). The relevant thermal current fluctuations S_{th} were calculated via $S_{\text{th}} = S_v G_{\text{H}}^2$, where $G_{\text{H}} = \nu e^2/h$. The voltage fluctuations were amplified by a cascade of a ‘cold’ (at 4.2 K) and a room-temperature amplifier and measured by a spectrum analyser. In our setup, the voltage gain of the cold amplifier, calibrated via thermal and shot-noise measurements (normalized to a bandwidth of 30 kHz), was about 9, with an input-referred noise of $260 \text{ pV Hz}^{-1/2}$. The gain of the room-temperature amplifier was 200, with an input-referred noise of $0.5 \text{ nV Hz}^{-1/2}$.

Following the procedure described in ref. ²⁶ (see Methods), the temperature T_m was plotted as a function of the dissipated power, ΔP . Additional contributions to the thermal conductance (for example, from phonons and bulk electrons) that depend only on T_m , were subtracted (hereafter referred to as ‘subtraction procedure’); namely, $\delta P_{\Delta N} = \Delta P(N_i, T_m) - \Delta P(N_j, T_m)$, where $i > j$, with $N_k = 2, 3, 4$ being the total number of open arms. This procedure allows a direct determination of the change of the heat flow due to changes in the number of conducting arms. Following this analysis, we plot a normalized coefficient for $\Delta N = N_i - N_j$ open arms, $\lambda_{\Delta N}/\Delta N = \delta P_{\Delta N}/(\kappa_0/2)$, as a function of T_m^2 , with the slope being the normalized thermal conductance, K/κ_0 , of a single arm (or a single mode). To strengthen our conclusions we also show selected data of (1) the total heat transport of N_k open arms, which contains also heat flux that is not associated with the chiral edge modes and (2) data analysis that includes the phonon contribution.

The molecular beam epitaxy (MBE)-grown GaAs-AlGaAs heterostructures that hosted the two-dimensional electron gas were designed to screen effectively the ionized dopants (Extended Data Fig. 1). Consequently, ‘relatively free’ electrons that reside in the doping regions can contribute to the electrical and thermal conductance. We carried out initial experiments with an extremely high-mobility two-dimensional electron gas by employing short-period superlattice doping (SPSL), with excess electrons in the doping layer²⁹. The low-temperature dark mobility was $31 \times 10^6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and the areal electron density was $3.1 \times 10^{11} \text{ cm}^{-2}$. Clear Hall plateaus and a longitudinal resistance of $R_{xx} < 10 \Omega$ were observed for filling factors of $\nu = 2-3$ (Extended Data Fig. 2). Although the electrons in the doping layer were relatively localized (owing to their heavy mass and the disorder), they still seemed to conduct heat. For example, when testing the heat flow at $\nu = 2$, where the thermal conductance is well understood, we found it to be higher than the expected value by approximately $3\kappa_0 T$. The subtraction procedure led to the expected heat conductance in each arm (Extended Data Fig. 3); yet, the accuracy required in these experiments necessitated a minimization of such unwanted contributions.

Consequently, we developed a ‘delta-doped’ scheme in a low-Al-mole-fraction $\text{Al}_x\text{Ga}_{1-x}\text{As}$ layer ($x \approx 0.23-0.25$). This led to shallower silicon deep-donor-like levels³⁰. As in the SPSL scheme, excess doping was required to obtain good conductance quantization in states with $\nu = 2-3$. The shallower donor level allowed a relatively low ‘freezing temperature’ of the electrons and thus a still efficient screening of the donors. However, sufficient charge localization was

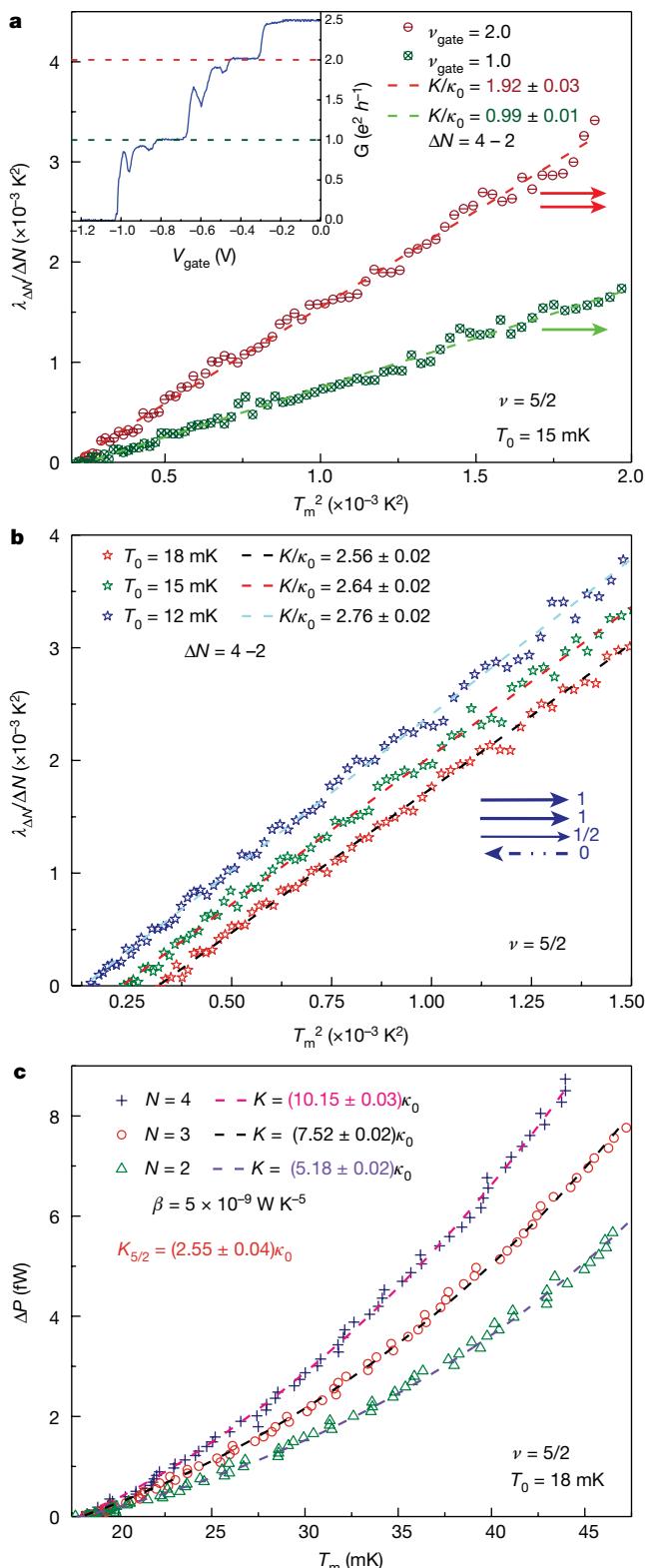


Fig. 2 | Heat flow of the two outermost edge modes at bulk filling

$\nu = 5/2$. **a**, The normalized heat flow $\lambda_{\Delta N}/\Delta N = \delta P_{\Delta N}/(\kappa_0/2)$, with δP being the difference in heat dissipation between $N = 4$ (four open arms) and $N = 2$ (two arms are open), with $\Delta N = 2$, as a function of T_m^2 . Two different realizations were studied: two downstream edge modes per arm (two red arrows, $\nu_{\text{gate}} = 2$) and one downstream edge mode per arm (one green arrow, $\nu_{\text{gate}} = 1$). The slopes of the fitted lines represent the normalized heat conductance, $K/\kappa_0 \approx 1.92$ for $\nu_{\text{gate}} = 2$ and $K/\kappa_0 \approx 0.99$ for $\nu_{\text{gate}} = 1$. The errors in the slopes are regressive errors (in the least-squares fitting procedure). The inset shows the transmission of a typical arm gate as a function of the gate voltage, with two plateaus that correspond to two or one edge modes propagating in the arm. **b**, Similar measurements with four and two arms fully transmitting (zero or slightly positive gate voltage). Measurements were performed at three different base electron temperatures, T_0 , with K/κ_0 increasing at lower temperatures. The arrows describe an idealized edge-mode structure of the particle-hole Pfaffian order as in Fig. 1a. **c**, Total power dissipation with $N = 4, 3$ and 2 , plotted as a function of T_m for $T_0 = 18 \text{ mK}$. K is determined by fitting the data with the expression $\Delta P = 0.5NK(T_m^2 - T_0^2) + \beta(T_m^5 - T_0^5)$ for the established phonon coefficient, β (see text). An average $K = 2.54\kappa_0$ is found. All of the errors mentioned here are confidence levels of between 95% and 99%.

achieved at base temperature, with negligible bulk heat conductance. We conducted measurements in a sample with a dark mobility of about $20 \times 10^6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and areal electron density $2.8 \times 10^{11} \text{ cm}^{-2}$, with a smaller span of the quantized Hall plateaus in a magnetic field and a higher lowest R_{xx} (Fig. 1b).

The sample was cycled three times between room temperature and low temperatures. This cycling process is known to change the distribution of charged impurities and donors, effectively rendering a slightly different device microscopically. Measurements were repeated at

different temperatures and different magnetic fields (in particular, at different fillings on the $\nu = 5/2$ plateau). Deducing the thermal conductance with a reasonable accuracy necessitates a careful determination of the parameters of the system. The following important factors were considered in the measurements (see also Methods). (i) Source noise. A non-ideal source contact may produce noise, which in turn will add to the measured thermal noise. Being uncorrelated with the thermal noise, if small enough, this noise is easily subtracted from the measured thermal noise. (ii) Equal division of currents and the overlap of Hall plateaus among the different arms (Extended Data Fig. 4). Although in general the uniformity of the electron density was excellent and the contact resistance was isotropic, small adjustments were done, when needed, by changing slightly the magnetic field. (iii) Because the small floating contact does not have zero contact resistance, incident (or emitted) currents may suffer reflection. Reflection was found to depend on the base temperature, the cycling round and the filling factor. In general, the reflection coefficient was found to be always smaller than 3%, with no remarkable difference between the thermal conductance determined in different runs. (iv) The stability of the electron base temperature, T_0 , during the long measurements of the thermal noise is crucial. Hence, consecutive measurements at different temperatures were performed after long intervals that allowed the electrons to reach a constant temperature. (v) An accurate determination of the amplification in the different Hall states under study is crucial. The amplification did not change considerably among the nearby $\nu = 7/3, 5/2$ and $8/3$ states studied (see Methods). (vi) Systematic errors resulted mostly from the uncertainty ($\pm 0.5 \text{ mK}$) in determining the electrons' temperature, which is directly tied to accuracy in determining the correct amplifier gain. We estimated the total error in the determination of the thermal conductance coefficient (due to regression and systematic errors) to be less than $\pm 0.05\kappa_0$ at a confidence level of more than 90%. Interestingly, the correct determination of T_0 led to an excellent linear-dependence fit of the power dissipation with T_m^2 (at low enough temperatures, where phonon contribution is negligible, or by employing the subtraction procedure) for any number N of open arms. (vii) During the measurements, the electrons initially remained at temperatures of around 18–20 mK (with a fridge temperature of about 10 mK) and later slowly cooled to 11–12 mK (with a fridge temperature $< 6 \text{ mK}$). Measurements at higher temperatures necessitated heating up the fridge.

Measurement results

We start by presenting the measurements performed at a bulk filling factor of $\nu = 5/2$. To verify the integrity of our measurements (for example, the amplification and the electron temperature), we first

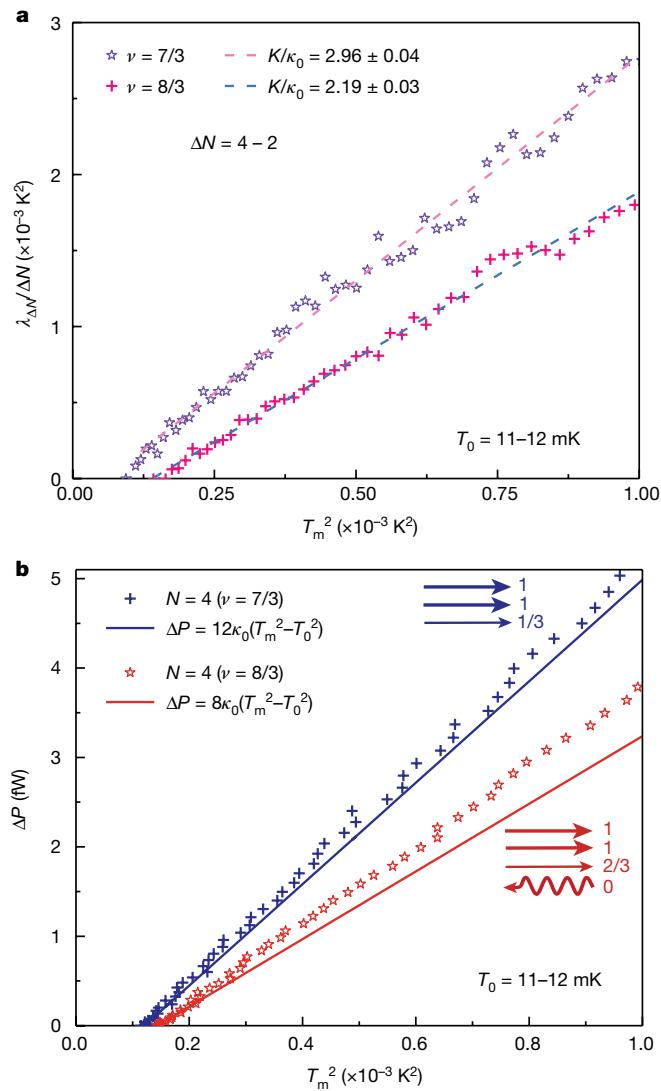


Fig. 3 | Normalized heat conductance at bulk fillings $\nu = 7/3$ and $\nu = 8/3$. **a**, A similar measurement to that shown in Fig. 2a for the two neighbouring fractional states to $\nu = 5/2$. For the $\nu = 7/3$ state $K_{7/3} = (2.96 \pm 0.05)\kappa_0$, with average $K_1 = (0.99 \pm 0.03)\kappa_0$ for a single edge mode. In the $\nu = 8/3$ state (errors mentioned are at a confidence level of better than 95%), with an upstream propagating neutral mode, $K = 2\kappa_0$ is expected for an infinite propagation length. The observed thermal conductance coefficient is larger (similar to that of the $\nu = 2/3$ state²⁶). **b**, The total dissipated power is plotted for the two states in **a** as a function of T_m^2 (without any data manipulation). Arrows describe the edge structure of each of the states. The measured power dissipation at $\nu = 7/3$ agrees well with the expected one, with a weak deviation at higher temperatures (likely due to the phonon contribution). In the $\nu = 8/3$ state the deviation is greater owing to the lack of equilibration, combined with bulk heat conductance due to the finite R_{xx} .

measured the heat conductance of the two outmost integer edge modes. As verified previously^{25,26}, the thermal conductance is expected to be an integer multiple of $\kappa_0 T$. By tuning the voltage of the arm gates (Fig. 1a), thus controlling the fillings under the gates, to either $\nu_{\text{gate}} = 1$ or $\nu_{\text{gate}} = 2$, the outermost-edge mode or the two outer-edge modes, respectively, were allowed to reach the remote grounds (inset, Fig. 2a). Under these conditions, the inner modes of the fractional $\nu = 1/2$ state were fully reflected back into the floating contact. Measurements were performed with $N = 4, 3$ and 2 , and here we plot the normalized electronic power dissipation of a single arm, $\lambda_{\Delta N}/\Delta N$ for $\Delta N = 2$, as a function of T_m^2 (following the subtraction procedure). The slope of each curve corresponds to the corresponding normalized heat conductance

coefficient, K/κ_0 (Fig. 2a). For $\nu_{\text{gate}} = 2$, the heat conductance for the two modes is $K_2 = (1.92 \pm 0.05)\kappa_0$, whereas for $\nu_{\text{gate}} = 1$ we find $K_1 = (0.99 \pm 0.04)\kappa_0$, with an average heat conductance per mode of $K_1 = (0.97 \pm 0.03)\kappa_0$. The small deviations from the expected values emanate from systematic and random errors (see above), as well as from possible non-ideality of the system. Yet, the results reproduce recent measurements of integer modes in altogether different devices²⁶.

In Fig. 2b we present the most important results of this work measured at the $\nu = 5/2$ state. With the gates unbiased, the heat was carried away from the floating reservoir into each open arm. Because this state may harbour counter-propagating modes, downstream charge modes and upstream neutral modes, a complete equilibration of the modes within each arm's edge is necessary to correctly determine the order of the $\nu = 5/2$ state. Hence, a propagation length substantially longer than the equilibration length (which is expected to decrease as the temperature increases²⁶) is required (here, $L \approx 150 \mu\text{m}$). A length that is too short may also harbour non-topological counter-propagating edge modes due to undesirable edge reconstruction. However, bulk contribution may have an important role for long arms because it may induce thermal flux backflow. We start by analysing the data with the subtraction procedure, which is applied to the data from the centre of the conductance plateau at three different base electron temperatures, T_0 . By plotting the normalized electronic power dissipation of a single arm, $\lambda_{\Delta N}/\Delta N$ for $\Delta N = 2$, as a function of T_m^2 , we observe a linear dependence at the three base temperatures of the electrons (confirming the correct determination of the temperature). As the temperature decreases, an increase in the heat conductance is evident, which we attribute to an increase of the temperature equilibration length. At higher temperatures, the thermal conductance tends to saturate with $K_{5/2} = (2.53 \pm 0.04)\kappa_0$, whereas at the lowest temperature we find $K_{5/2} = (2.76 \pm 0.04)\kappa_0$; we address this observation in more detail later.

To validate our measurement results, we present also raw data in Fig. 2c (and Fig. 3). In the plot, the total power dissipation ΔP at $T_0 = 18 \text{ mK}$ is shown as a function of T_m for different numbers of arms. We find $K_{5/2} = (2.55 \pm 0.04)\kappa_0$ for a single, fully open arm. The dashed lines are least-squares fits using $K(T_m^2 - T_0^2) + \beta(T_m^5 - T_0^5)$, with $\beta = 5 \times 10^{-9} \text{ W K}^{-5}$, where K is a single free parameter used to fit the data points. Fixing a single β (which depends on the size of our floating contact) as above leads to an uncertainty in K of $\sim 0.02\kappa_0$ in the present analysis; this agrees with the uncertainty in the subtraction procedure. As can be seen in the figures, the fits across the full temperature range are in excellent agreement with the data points, as the chosen β provides the lowest accumulated error in the fitting line (in all our datasets).

We then studied the two neighbouring fractional states, the simpler $\nu = 7/3$ state and the $\nu = 8/3$ hole-conjugate state. The $\nu = 7/3$ state is expected to support only three downstream charge modes: two integer and one fractional (the inner $\nu = 1/3$, $e^* = e/3$)³¹, each with central charge²⁶ equal to 1. The hole-conjugate $\nu = 8/3$ state, similarly to the $\nu = 2/3$ state in the lowest Landau level, also supports three downstream charge modes, two integer and one fractional (the inner $\nu = 2/3$, with $e^* = e/3$)³¹, but the inner mode is accompanied by an upstream neutral mode (with central charge equal to 1). Effectively, in fully equilibrated transport, a net of two modes carries the heat²⁶.

Representative results in the above two states, obtained using the subtraction procedure at $T_0 = 11-12 \text{ mK}$, are shown in Fig. 3b (see also Extended Data Fig. 5). In Fig. 3a, for $T_m(\text{max}) \approx 30 \text{ mK}$, $\nu = 7/3$ and for a fully open arm we find the thermal conductance coefficient to be $K_{7/3} = (2.96 \pm 0.04)\kappa_0$. This corresponds to an average of $K_1 = (0.99 \pm 0.03)\kappa_0$ per single charge edge mode (integer or fractional). For the $\nu = 8/3$ state, we find a thermal conductance coefficient of $K_{8/3} = (2.19 \pm 0.03)\kappa_0$ instead of the expected $K_{8/3} = 2\kappa_0$ for fully equilibrated counter-propagating modes. We note also the finite R_{xx} at $\nu = 8/3$ (larger than in the other two neighbouring states; see Fig. 1b), which may lead to heat leakage into the bulk, thus effectively increasing the apparent thermal conductance. It is important to recall that in previous measurements of the heat conductance at the fractional $\nu = 2/3$ state we found $K_{2/3} = 0.33\kappa_0$ at $T_0 = 10 \text{ mK}$ and $K_{2/3} = 0.25\kappa_0$ at

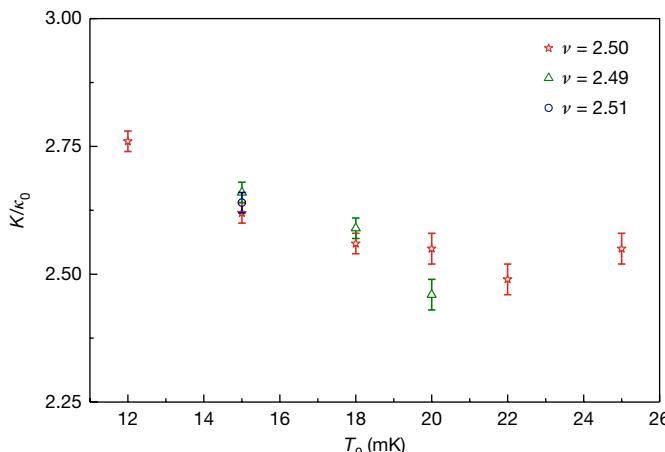


Fig. 4 | Summary of the normalized thermal conductance coefficient results for $\nu = 5/2$. Plotted is the average K/κ_0 as a function of the temperature at three different fillings on the $\nu = 5/2$ G_H conductance plateau. A clear tendency of increased thermal conductance at lower temperatures is visible. Such dependence is attributed to the increased equilibration length (among downstream and upstream modes) at lower temperatures (see ref. ²⁶ for a similar behaviour of the $\nu = 2/3$ state). Seventeen measurements were conducted, with K/κ_0 falling in the range $K/\kappa_0 = (2.53 \pm 0.04)\kappa_0$ at electron base temperatures $T_0 = 18\text{--}25$ mK, where most of the data points were taken.

$T_0 = 30$ mK. Such dependence on the temperature is expected because it is tied directly with the temperature dependence of the equilibration length (see methods in ref. ²⁶). The presence of neutral modes in $\nu = 8/3$ and in $\nu = 5/2$ was also verified (see Extended Data Fig. 6).

In Fig. 3 we show raw data for the two neighbouring states near $\nu = 5/2$ without any data manipulation (such as the subtraction procedure or taking into account the phonon contribution). The total heat dissipation for $N = 4$ is plotted in Fig. 3b as a function of T_m^2 . For the $\nu = 7/3$ state, a deviation from $K_1 = \kappa_0$ for a single charge mode (namely, $K_{7/3} = 3\kappa_0$ in each arm) becomes evident when the temperature approaches about 30 mK, which is clearly due to the phonon contribution. Yet, for the $\nu = 8/3$ state, the deviation from the expected total thermal conductance of $8\kappa_0$ (for four arms) is more apparent. Here, the non-equilibrated heat transport, combined with additional bulk transport (because $R_{xx} > 0$), contributes to a larger deviation of the thermal conductance.

In Fig. 4 we summarize the normalized thermal conductance coefficient of the $\nu = 5/2$ state. We plot K/κ_0 at different temperatures (12–25 mK) and at three filling factors (on the $\nu = 5/2$ conductance plateau). Seventeen measurements were performed with the device temperature cycled three times to room temperature, thus allowing for different microscopic configurations of the ionized charges. A saturation of the normalized thermal conductance $K_{5/2} = (2.53 \pm 0.04)\kappa_0$ is observed in the temperature range $T_0 = 18\text{--}25$ mK. The thermal conductance increases when the temperature is ≤ 15 mK. Such dependence is attributed to the increased equilibration length among the counter-propagating modes at lower temperatures (see the next section and Methods). In Fig. 5 we have summarized some of the more likely edge structures of possible orders that may describe the $\nu = 5/2$ state, as well as their equilibrated K/κ_0 values.

Discussion and conclusions

Composite fermions and the K-matrix formalism provide a powerful framework for the understanding of the fractional quantum Hall effect in the lowest Landau level. Almost all the quantum Hall states in the lowest Landau level are believed to be integer quantum Hall liquids of composite fermions³². Our recent results on thermal transport in the lowest Landau level strongly support that picture²⁶. The first excited Landau level poses a greater challenge. The nature of the $\nu = 5/2$ and $\nu = 12/5$ states has long been a puzzle. Competing proposals have also

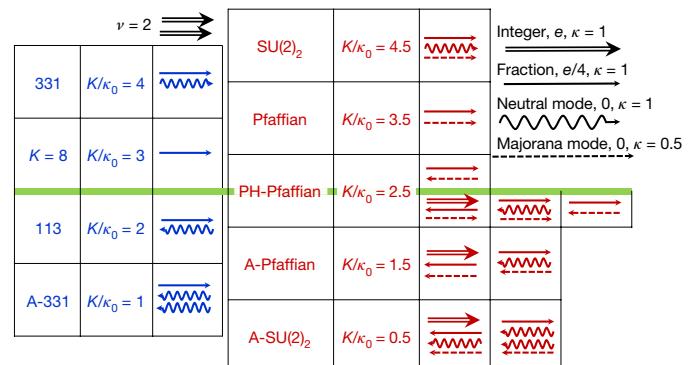


Fig. 5 | Possible orders predicted for the $\nu = 5/2$ state. Edge-mode structure of the leading candidates for the many-body state of a fractional quantum Hall $\nu = 5/2$ liquid: Pfaffian, anti-Pfaffian (A-Pfaffian) and particle-hole Pfaffian (PH-Pfaffian) topological orders and the $SU(2)_2$, $K = 8$, 331 and 113 liquids ('A' stands for 'anti'). Their expected quantized thermal Hall conductance, K_T , in units of $\kappa_0 T$ are also shown. A right-pointing double-line arrow denotes a downstream edge mode of a fermion with charge $e^* = e$, contributing Hall conductivity $G_H = e^2/h$ and $K/\kappa_0 = 1$. Right- and left-pointing solid-line arrows denote a downstream and an upstream fractional charge mode, respectively, contributing $0.5G_H = e^2/(2h)$ and $K/\kappa_0 = 1$. The wavy line denotes a fermionic neutral mode with zero charge and $K/\kappa_0 = 1$, and the dashed line denotes a Majorana mode with zero charge and $K/\kappa_0 = 1/2$. A neutral mode with $K/\kappa_0 = 1$ is physically equivalent to two Majorana modes. The left (right) part of the figure depicts the Abelian (non-Abelian) states with an integer (half-integer) K/κ_0 . For all states, the lowest Landau level is fully occupied, as shown by the two downstream fermions (the two black right-pointing double-line arrows in the top left part of the figure). The $\nu = 5/2$ state can be constructed in a particle-like manner, starting from $\nu = 2$ and adding fractional, neutral, or Majorana modes, or in a hole-like fashion, starting from $\nu = 3$ and adding modes moving in opposite directions (for example, see the third column of the anti-Pfaffian phase). The final state of the edge modes, after full equilibration, is depicted in the right-most column of each row. The green line divides the particle-like and hole-like states.

been made for the $\nu = 7/3$ and $\nu = 8/3$ states^{33,34}; however, past^{31,35} and present results prove that these two states are compatible with composite fermion liquids, similar to those at $\nu = 1/3$ and $\nu = 2/3$.

We now discuss our measurement results and their implications. In this study and in our previous work²⁶, the results agree well with theoretical predictions³⁶, with a level of precision of about 1% (in sufficiently long samples). Two of the states require further elaboration. For the $\nu = 2/3$ state, theory predicts a thermal Hall conductance that scales like $1/L$ (with L being the system size), with a proportionality constant that decreases with temperature (see methods in ref. ²⁶). Our measurements are compatible with this expectation. For the $\nu = 8/3$ state, we observed a thermal Hall conductance exceeding the expected one by about 10%, which we also attribute to partial equilibration of counter-propagating edge modes and the finite R_{xx} of this state. When examining the case of $\nu = 5/2$, these results give confidence in our experimental setup and analysis.

For the $\nu = 5/2$ state, taking our findings of $K_{5/2} = 2.5\kappa_0$ at face value, the results are compatible with the topological order of the particle-hole Pfaffian liquid^{18,19}. On the basis of common theoretical understanding, the fractional value of K implies that the $\nu = 5/2$ state is non-Abelian. The identification of this topological order is rather surprising, in view of the numerical works that predicted the Pfaffian and anti-Pfaffian topological orders as leading candidates. This discrepancy may be reconciled by models that include disorder^{37–39}.

There may be factors that would cause the measured thermal Hall conductance to be different from that imposed by the bulk topological order. Of these, we note the lack of equilibration of counter-propagating edge modes and leakage of heat to the bulk either due to longitudinal electronic thermal conductance or due to phonons. We were able to quantify the phonons' contribution and ascertain that it is small

(otherwise, we fully subtracted it). The observation of plateau-like saturation in the measured two-terminal thermal conductance suggests that the contribution of the longitudinal conductance is small, although the level of precision of the plateau cannot rule out that such contribution exists.

To analyse possible effects of partial equilibration, we consider an edge carrying upstream modes with a thermal Hall conductance coefficient K_u and downstream modes with a thermal Hall conductance coefficient K_d . Full equilibration at the edge, which is expected for long enough samples, would result in a measured thermal Hall conductance coefficient of $|K_d - K_u|$, whereas in the absence of any equilibration the measured thermal Hall conductance coefficient would be $K_d + K_u$. Partial equilibration may give rise to intermediate values between these two limits. If some of the modes on an edge fully equilibrate while others stay intact, a plateau will arise with a heat conductance that is different from that dictated by the topological order in the bulk (with a possible distribution of the heat current between the two edges of the sample). Remarkably, the difference between the values of the heat conductance in the two models is an integer. Thus, an observed half-integer value in this case implies that the topological order is non-Abelian. In particular, such a picture (with equilibration of only certain modes) might be consistent with the anti-Pfaffian topological order⁴⁰. Hence, although our results point at the $\nu = 5/2$ state being non-Abelian, they might not fully identify its topological order.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0184-1>.

Received: 30 September 2017; Accepted: 26 March 2018;

Published online 4 June 2018.

1. Nayak, C. et al. Non-Abelian anyons and topological quantum computation. *Rev. Mod. Phys.* **80**, 1083–1159 (2008).
2. Willett, R. et al. Observation of an even-denominator quantum number in the fractional quantum Hall effect. *Phys. Rev. Lett.* **59**, 1776–1779 (1987).
3. Moore, G. et al. Nonabelions in the fractional quantum Hall effect. *Nucl. Phys. B* **360**, 362–396 (1991).
4. Greiter, M. et al. Paired Hall state at half filling. *Phys. Rev. Lett.* **66**, 3205–3208 (1991).
5. Read, N. et al. Paired states of fermions in two dimensions with breaking of parity and time-reversal symmetries and the fractional quantum Hall effect. *Phys. Rev. B* **61**, 10267–10297 (2000).
6. Dolev, M. et al. Observation of quarter of an electron charge at the $\nu = 5/2$ quantum Hall state. *Nature* **452**, 829–834 (2008).
7. Radu, I. P. et al. Quasi-particle properties from tunneling in the $\nu = 5/2$ fractional quantum Hall state. *Science* **320**, 899–902 (2008).
8. Bid, A. et al. Observation of neutral modes in the fractional quantum Hall regime. *Nature* **466**, 585–590 (2010).
9. Morf, R. H. Transition from quantum Hall to compressible states in the second Landau level: new light on the $\nu = 5/2$ enigma. *Phys. Rev. Lett.* **80**, 1505–1508 (1998).
10. Storni, M. et al. Fractional quantum Hall state at $\nu = 5/2$ and the Moore–Read Pfaffian. *Phys. Rev. Lett.* **104**, 076803 (2010).
11. Rezayi, E. H. Landau level mixing and the ground state of the $\nu = 5/2$ quantum Hall effect. *Phys. Rev. Lett.* **119**, 026801 (2017).
12. Levin, M. et al. Particle-hole symmetry and the Pfaffian state. *Phys. Rev. Lett.* **99**, 236806 (2007).
13. Lee, S. S. et al. Particle-hole symmetry and the $\nu = 5/2$ quantum Hall state. *Phys. Rev. Lett.* **99**, 236807 (2007).
14. Wen, X. G. Non-Abelian statistics in the fractional quantum Hall states. *Phys. Rev. Lett.* **66**, 802–805 (1991).
15. Halperin, B. I. Theory of the quantized Hall conductance. *Helv. Phys. Acta* **56**, 75–102 (1983).
16. Yang, G. et al. Influence of device geometry on tunneling in $\nu = 5/2$ quantum Hall liquid. *Phys. Rev. B* **88**, 085317 (2013).
17. Yang, G. et al. Experimental constraints and a possible quantum Hall state at $\nu = 5/2$. *Phys. Rev. B* **90**, 161306 (2014).
18. Son, D. T. Is the composite fermion a Dirac particle? *Phys. Rev. X* **5**, 031027 (2015).
19. Zucker, P. T. et al. Stabilization of the particle-hole Pfaffian order by Landau-level mixing and impurities that break particle-hole symmetry. *Phys. Rev. Lett.* **117**, 096802 (2016).

20. Fidkowski, L. et al. Non-Abelian topological order on the surface of a 3D topological superconductor from an exactly solved model. *Phys. Rev. X* **3**, 041016 (2013).
21. Bonderson, P. et al. A time-reversal invariant topological phase at the surface of a 3D topological insulator. *J. Stat. Mech.* **2013**, P09016 (2013).
22. Kane, C. L. et al. Pairing in Luttinger liquids and quantum Hall states. *Phys. Rev. X* **7**, 031009 (2017).
23. Schwab, K. et al. Measurement of the quantum of thermal conductance. *Nature* **404**, 974–977 (2000).
24. Meschke, M. et al. Single-mode heat conduction by photons. *Nature* **444**, 187–190 (2006).
25. Jezouin, S. et al. Quantum limit of heat flow across a single electronic channel. *Science* **342**, 601–604 (2013).
26. Banerjee, M. et al. Observed quantization of anionic heat flow. *Nature* **545**, 75–79 (2017).
27. Wen, X. G. *Quantum Field Theory of Many-body Systems: From the Origin of Sound to an Origin of Light and Electrons* (Oxford Univ. Press, Oxford, 2004).
28. Wellstood, F. C. et al. Hot-electron effects in metals. *Phys. Rev. B* **49**, 5942–5955 (1994).
29. Umansky, V. et al. in *Molecular Beam Epitaxy: From Research to Mass Production* (ed. Henini, M.) 121–137 (Elsevier, Amsterdam, 2013).
30. Mooney, P. M. Deep donor levels (DX centers) in III–V semiconductors. *J. Appl. Phys.* **67**, R1–R26 (1990).
31. Dolev, M. et al. Characterizing neutral modes of fractional states in the second Landau level. *Phys. Rev. Lett.* **107**, 036805 (2011).
32. Jain, J. K. *Composite Fermions* (Cambridge Univ. Press, Cambridge, 2007).
33. Read, N. et al. Beyond paired quantum Hall states: parafermions and incompressible states in the first excited Landau level. *Phys. Rev. B* **59**, 8084–8092 (1999).
34. Bonderson, P. et al. Fractional quantum Hall hierarchy and the second Landau level. *Phys. Rev. B* **78**, 125323 (2008).
35. Dolev, M. et al. Dependence of the tunneling quasiparticle charge determined via shot noise measurements on the tunneling barrier and energetics. *Phys. Rev. B* **81**, 161303 (2010).
36. Kane, C. L. et al. Quantized thermal transport in the fractional quantum Hall effect. *Phys. Rev. B* **55**, 15832–15837 (1997).
37. Mross, D. F. et al. Theory of disorder-induced half-integer thermal Hall conductance. Preprint at <https://arxiv.org/abs/1711.06278> (2017).
38. Wang, C., Vishwanath, A. & Halperin, B. I. Topological order from disorder and the quantized Hall thermal metal: possible applications to the $\nu = 5/2$ state. Preprint at <https://arxiv.org/abs/1711.11557> (2017).
39. Lian, B. et al. Theory of disordered $\nu = 5/2$ quantum thermal Hall state: emergent symmetry and phase diagram. *Phys. Rev. B* **97**, 165124 (2018).
40. Steven, S. H. On the interpretation of thermal conductance of the $\nu = 5/2$ edge. *Phys. Rev. B* **97**, 121406 (2018).

Acknowledgements We acknowledge B. Halperin and S. Simon for discussions. M.B. acknowledges the help and advice of Y. Gross regarding fabrication processes and R. Bhattacharya for help with the cold amplifiers and Y. C. Chung and H. K. Choi for their help with the dilution refrigerator. M.H. acknowledges the continuous support of the Sub-Micron Center staff, and in particular Y. Rotblat, without whom this work would not be possible. M.H. acknowledges the support of the European Research Council under the European Community's Seventh Framework Program (FP7/2007–2013)/ERC under grant agreement number 339070, the partial support of the Minerva foundation under grant number 711752, the Israeli Science Foundation ISF under grant number 459/16 and, together with V.U., the German Israeli Foundation (GIF) under grant number I-1241-303.10/2014. A.S. and Y.O. acknowledge support from the European Research Council under the European Union's Seventh Framework Program (FP7/2007–2013)/ERC Project MUNATOP, the DFG (CRC/Transregi 183, El 519/7-1) and the Israel Science Foundation. Y.O. acknowledges the Binational Science Foundation (BSF). D.E.F.'s research was supported in part by the National Science Foundation under grant number DMR-1607451.

Reviewer information *Nature* thanks K. Shtengel, S. Simon and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.B. and M.H. designed the experiment, preformed the measurements, did the analysis and guided the experimental work. M.B. fabricated the devices with input from M.H., D.E.F. and Y.O., and A.S. worked on the theoretical aspects. V.U. grew the two-dimensional electron-gas heterostructures. All authors contributed to the write up of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0184-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

MBE-grown heterostructures. The MBE growth techniques used for heterostructures harbouring the fragile $\nu = 5/2$ fractional state were developed during the past decades. Poor correlation was found between the zero-field mobility and the $\nu = 5/2$ energy gap. The key factor influencing the robustness of the state was found to be the spatial correlations among the charged scatters in the doped layers²⁹. The highest quality (largest gap) was observed when using the SPSL scheme accompanied by controlled illumination at low temperatures^{41–43}. However, the fabricated devices exhibited poor temporal stability, as well as gate hysteresis⁴⁴. In this work, we used two types of structures, one that had substantial thermal bulk conductance and another that did not (see Extended Data Fig. 1).

Device fabrication. A high-purity heterojunction hosting a high-mobility two-dimensional electron gas, with areal electron density $2.8 \times 10^{11} \text{ cm}^{-2}$ and 4.2-K ‘dark’ mobility $\mu = 20 \times 10^6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, was used. The depth of the two-dimensional electron gas below the surface was 176 nm, the spacer layer (separation between two-dimensional electron gas and donors) was 85 nm and the quantum-well width was 30 nm. A 5-nm-thick HfO_2 layer was deposited on the surface using an atomic layer deposition process and served as a dielectric layer under the gates. Because the crystal direction is found to affect the contact resistance^{44,45}, ‘zigzag edge’-type contacts were fabricated. Etched grooves under the floating contact (not visible in the SEM micrograph) ensure that the incident current enters the bulk of the metal ohmic contact before splitting between the different arms of the device. Following a thorough cleaning of the surface (by plasma ashing and oxide removal), contacts were evaporated in an electron-gun evaporator with a base pressure of 10^{-8} torr. The evaporation sequence for ohmic contacts, from the GaAs surface and up, was: Ni (5 nm), Au (200 nm), Ge (100 nm), Ni (75 nm) and Au (15 nm). Contacts were alloyed at 450°C for 2 min. The continuous gates were evaporated on HfO_2 in the sequence Ti (5 nm) and Au (20 nm).

Continuous gates versus quantum point contact constrictions. Because the heterostructures are doped using the SPSL method²⁹, ‘relatively free’ electrons exist in the donor layer. Depleting, or pinching, electrons in the two-dimensional electron gas by using a quantum point contact (QPC) constriction requires a large negative gate voltage of about -10 V (in order to deplete also the donor layer). This leads to considerable hysteresis and instability, presumably due to the ‘slow’ motion of the carriers in the donor layers. By replacing the QPCs with continuous gates, deposited on a high-dielectric-constant HfO_2 insulator, the required voltage for depletion was about -1 V, thus minimizing the instability and hysteresis effects.

Small floating contact. Owing to the finite contact resistance, backscattering from the floating heated reservoir leads to an effective series thermal resistance and possibly shot noise in the source arm. Backscattering from the contact was measured by comparing the incident current to the reflected one when the number of the fully open arms was changed. The source current was initially fully reflected into the amplifier, thus measuring the impinging current, I_s . Then, in a two-arm configuration, the reflected current, I_{ref} was measured again. The reflection coefficient r was calculated from $r = 2I_{\text{ref}}/I_s - 1$. Similar expressions are used for any N open arms. The measured reflection depended on the cycling round and the filling factor and was always less than 3%.

Branching of current into N arms. Because the small floating contact might not have the same contact resistance in each of the four arms, and the density in each arm may be slightly different, the equal branching of the currents in each arm must be verified. Optimization was performed by a changing slightly the magnetic field (Extended Data Fig. 4).

Calibration of the gain and T_0 . Knowing the gain of the amplification chain is crucial for the determination of the electron temperature, T_m . Two calibration regions (for the two amplification chains), each composed of an additional gated region and two contacts (source and ground) were added in two opposite arms²⁶. Two methods were used to calibrate the gain: (a) verifying the well known quasiparticle charge at a known temperature and (b) measuring thermal noise when the electron temperature was equal to the fridge temperature (at temperatures higher than 70–100 mK). The different effective gains (at 30 kHz bandwidth) were determined at different filling factors by comparing the areas under the resonance curves.

For example, calibrating the gain (of the amplification chain, composed of two amplifiers and a spectrum analyser) via shot-noise measurements proceeded as follows. First, the temperature was determined, independently of the gain, by a linear extrapolation of the noise curve versus the current I (at $eV_s \gg 2k_B T_0$) to zero noise. The intersection point is $eV_s = 2k_B T_0$, with $V_s = Ih/(ve^2)$ being the Hall voltage (see Extended Data Fig. 5b and ref. ²⁶). With the transmission coefficient of the QPC, t_{QPC} , and the known quasiparticle charge, e^* , the gain is determined by a simple matching procedure of the shot noise data with the familiar expression of the spectral density of the noise. We use the usual expression of shot noise for independent scattering events, $S_s = 2e^* I t_{\text{QPC}} (1 - t_{\text{QPC}}) \zeta(V, T)$, where $\zeta(V, T)$ is a temperature-dependent factor that has been proven to work quite well in all the quantum Hall effect regimes and V is the applied voltage. T_0 is determined with an accuracy of ± 0.5 mK.

The gain is actually an ‘effective gain’ that depends on the bandwidth of the LCR_H circuit, with $R_H = 1/G_H$, and the spectrum analyser bandwidth. The gain

squared is proportional to the area under the Lorentzian power response. Hence, the ratio of these areas at different filling factors should equal the ratio of the squares of the gains. This method was compared with measurements of the thermal noise at elevated temperatures.

It is difficult to determine the systematic errors in the determination of the gain. However, we checked the effect of a slight modification of the gain on the determined temperature T_m and found it to be negligible in comparison with the scattering of the random errors seen in the figures. Interestingly, a correct determination of T_0 led to an extremely good functional fit of the power dissipation with $T_m^2 - T_0^2$ for any number N of open arms and filling factors.

Determining T_m . In a general case of multiple one-dimensional edge modes in each of the N arms, the expressions for the dissipated power and the noise are cumbersome. We express the dissipated power in the floating contact as

$$\Delta P = \frac{1}{2} \frac{I_s^2}{G_H} \frac{\nu_{\text{QPC1}}}{\nu^2} \left[1 - \frac{\nu_{\text{QPC1}}}{\sum_{i=1}^{i=4} \nu_{\text{QPCi}}} \right]$$

where ν_{QPCi} is the filling factor in the QPC. In turn, the temperature T_m is related to the thermal noise via $S_{\text{th}} = 2G^* k_B (T_m - T_0)$, with

$$\frac{1}{G^*} = \frac{1}{G_{\text{amp}}} + \frac{1}{\sum_{i=1, i \neq \text{amp}}^{i=4} G_i}$$

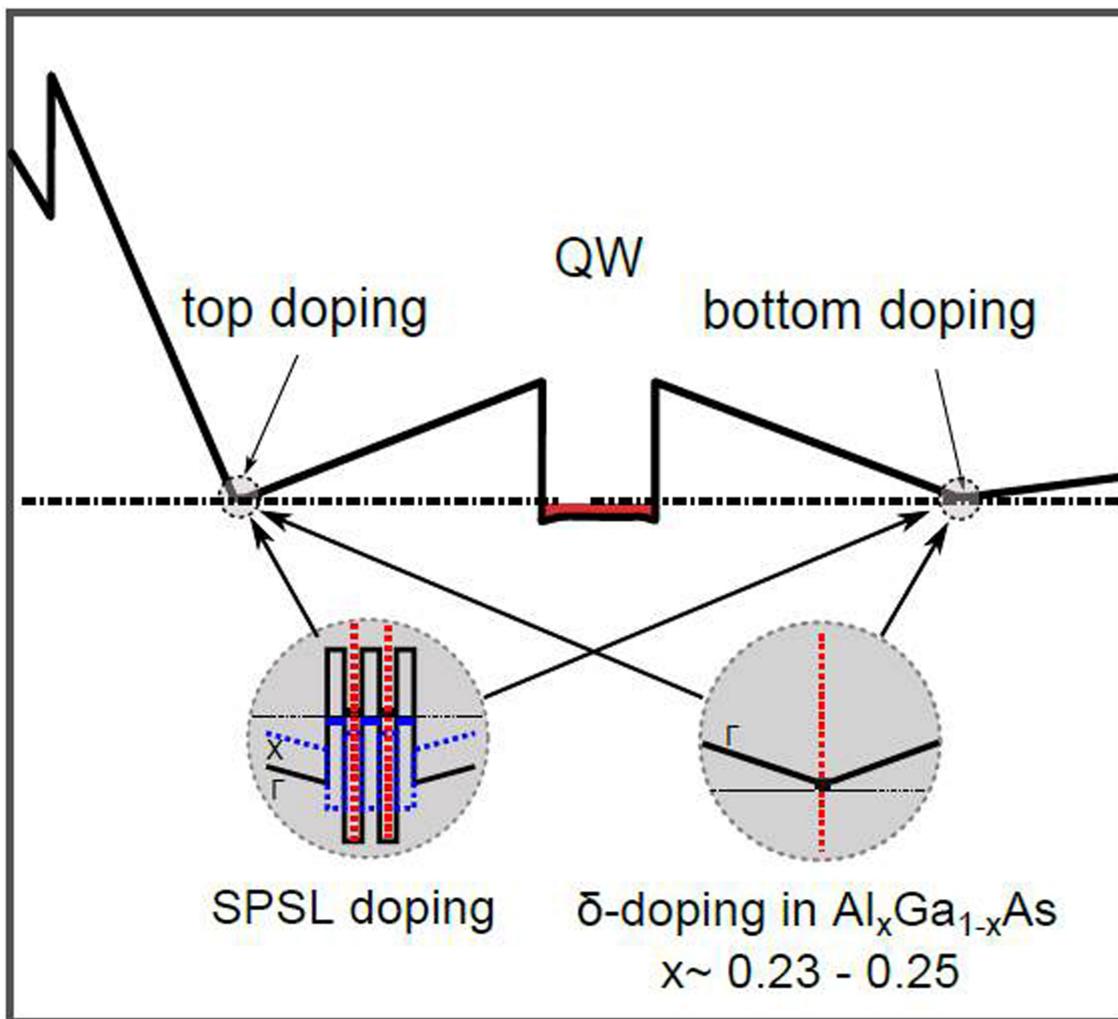
where G_i is the conductance of the i th arm^{23–26} and G_{amp} is the conductance of the arm that hosts the amplifier.

Thermal conductance at low temperatures. Here we discuss the observed increase in the thermal conductance at 12 mK at the $\nu = 5/2$ state that is due to the lack of equilibration between the upstream and downstream modes. We show that this behaviour is compatible with the particle–hole Pfaffian model. We argue that because the particle–hole Pfaffian model has only one upstream Majorana mode, the equilibration length $\xi(T)$ is especially long compared to all other quantum Hall states at the lowest temperatures. Thus, the upstream and downstream modes in the particle–hole Pfaffian state may require a higher temperature for equilibration compared to other topological liquids. In what follows, we only focus on the energy exchange between the upstream and downstream modes and do not address how local quasi-equilibrium is established in each mode. We also neglect tunnelling from the fractional channel to the integer ones. We believe that this is justified by the width of the confining potential of the etch-defined edge⁴⁶ and by spin conservation.

At the $\nu = 2/3$ state, the thermal conductances of upstream and downstream modes are equal. This leads to a relatively slow temperature dependence of the observed conductances as $K \approx \xi(T)/L$, where L is the edge length and $\xi(T)$ is the equilibration length²⁶. A zero thermal conductance is expected only at an infinite L . The exact dependence of $\xi(T)$ on the temperature is determined by the system details but we expect that $\xi(T)$ grows at low temperatures, in agreement with the data. At filling factors such as $\nu = 3/5$, the total thermal conductances of upstream and downstream modes differ. Then, the correction to the universal heat conductance varies as $\exp[-CL/\xi(T)]$, where C is a constant²⁶. Thus, only small corrections to the universal value are expected as long as $\xi(T) \ll L$. When $\xi(T) \geq L$, the thermal conductance should grow with decreasing temperature. Such growth has been seen experimentally only at $\nu = 5/2$ and $T < 15$ mK. This suggests that the inequality $\xi(T) \ll L$ holds at accessible temperatures for all states except $\nu = 5/2$ (ref. ³⁶).

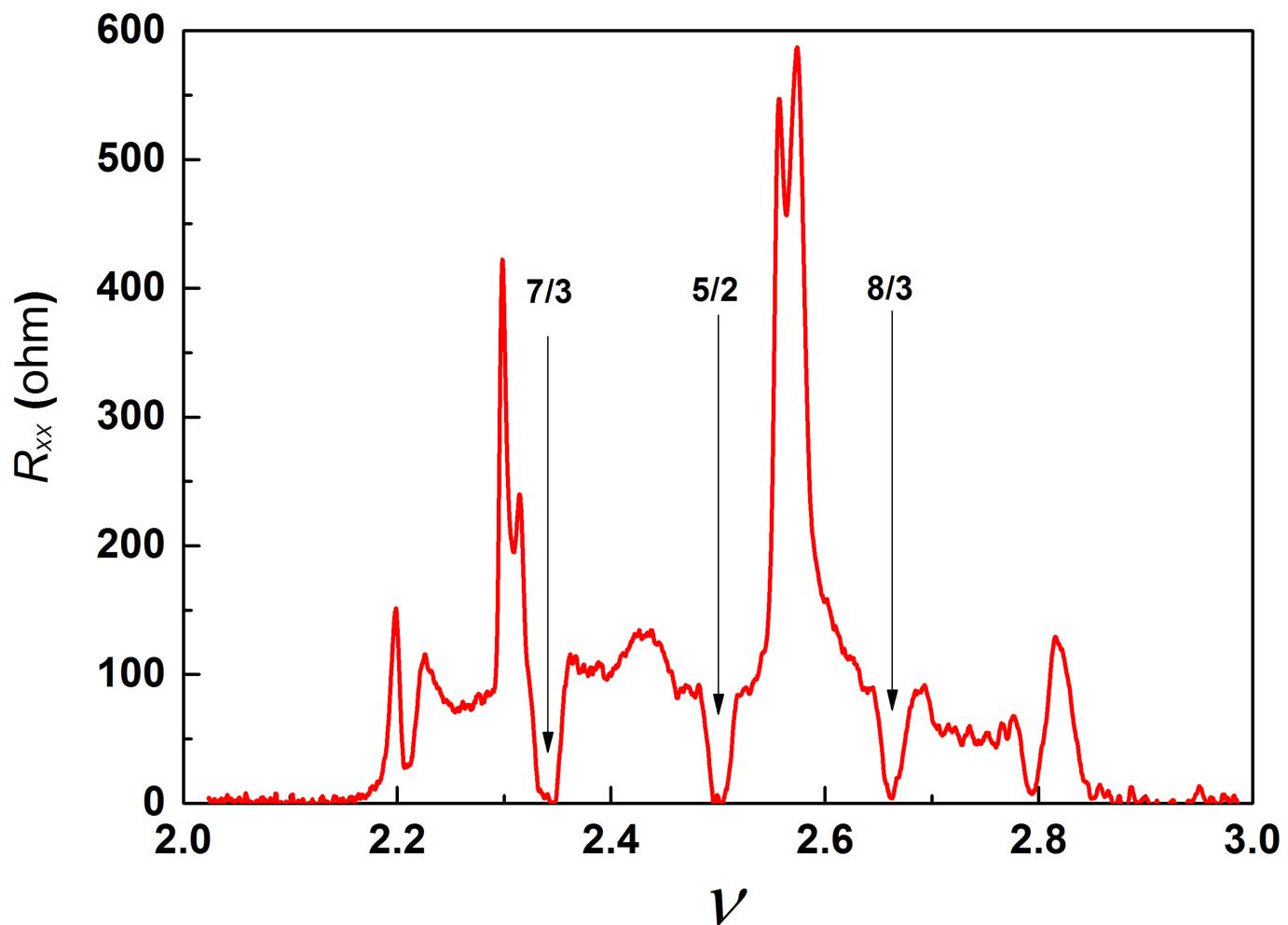
The different behaviour of $\xi(T)$ for different filling factors may originate from the structure of the most relevant inter-mode interaction. In Abelian states this interaction is $\nu(x) \partial_x \phi_i \partial_x \phi_j$, where ϕ_k denotes the k th Bose mode and the interaction amplitude $\nu(x)$ is a random function of coordinates. A state with at least two upstream Majorana modes ε_i exhibits the same scaling of the equilibration length because of the same scaling dimension of the leading inter-mode interaction $\nu(x) \varepsilon_i \varepsilon_j \partial_x \phi_k$. In the particle–hole Pfaffian model, there is only a single upstream Majorana mode ε , and therefore such a term cannot exist in the Hamiltonian. Thus, the interaction is less relevant in the renormalization group sense. This translates into a weaker thermal exchange at low temperatures and a faster growth of the equilibration length at low T .

41. Willett, R. L. The quantum Hall effect at 5/2 filling factor. *Rep. Prog. Phys.* **76**, 076501 (2013).
42. Samani, M. et al. Low-temperature illumination and annealing of ultrahigh quality quantum wells. *Phys. Rev. B* **90**, 121405 (2014).
43. Rössler, C. et al. Gating of high-mobility two-dimensional electron gases in GaAs/AlGaAs heterostructures. *New J. Phys.* **12**, 043007 (2010).
44. Slobodeniuk, A. O. et al. Equilibration of quantum Hall edge states by an Ohmic contact. *Phys. Rev. B* **88**, 165307 (2013).
45. Dahlem, F. Cryogenic scanning force microscopy of quantum Hall samples: adiabatic transport originating in anisotropic depletion at contact interfaces. *Phys. Rev. B* **82**, 121305 (2010).
46. Gelfand, B. Y. et al. Edge electrostatics and a mesa-etched sample and edge-state-to-bulk scattering rate in the fractional quantum Hall regime. *Phys. Rev. B* **49**, 1862–1866 (1994).



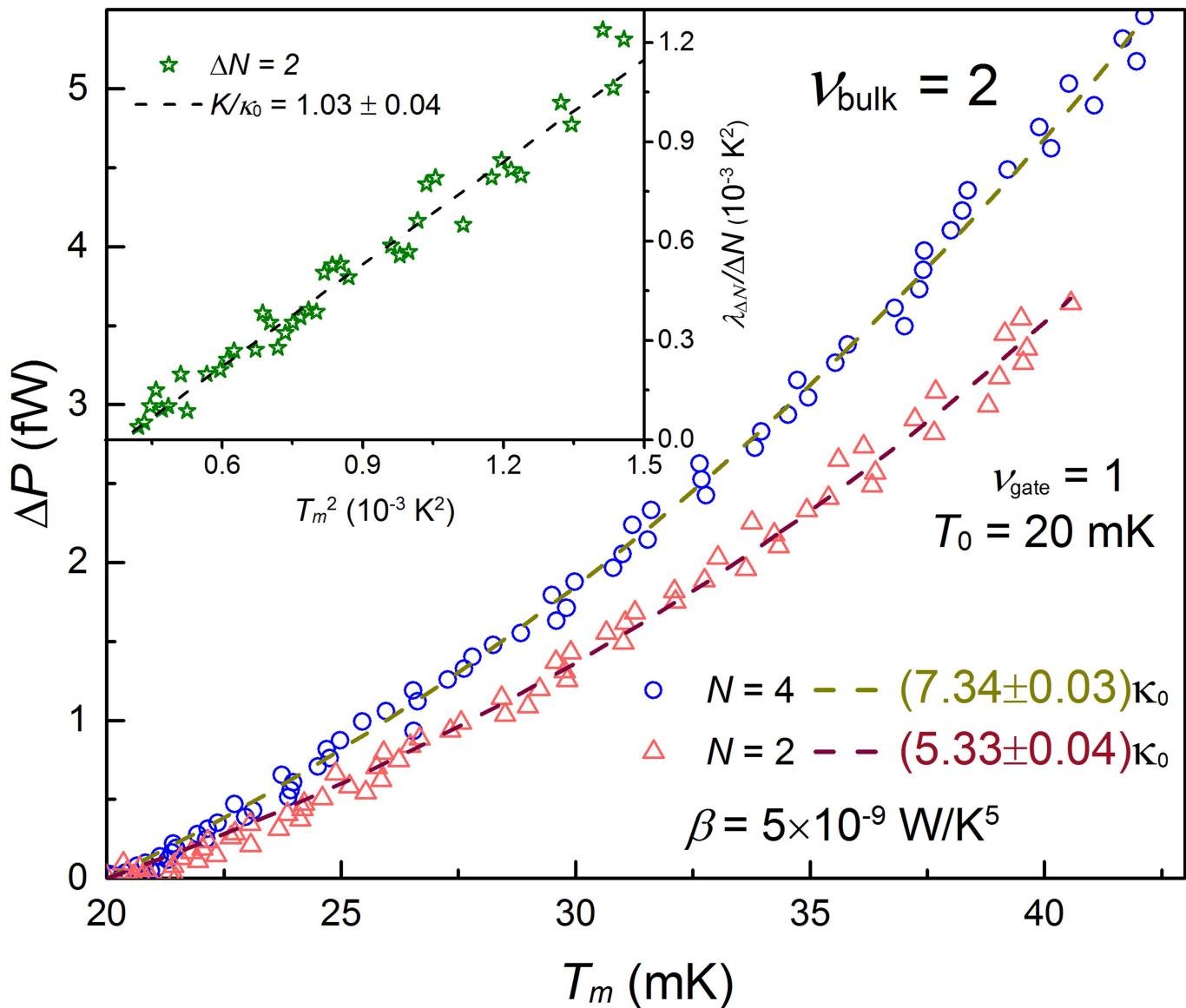
Extended Data Fig. 1 | Details of the growth structure. Schematic of the conduction band in the MBE-grown structures that were studied. The SPSL doping scheme comprises δ -Si doping planes placed in narrow GaAs quantum wells (QW). The thickness of the GaAs and AlAs quantum wells in SPSL is chosen in such a way that the X-band minima of the AlAs layers reside below the Γ -band minimum of the GaAs. Electrons that spill over

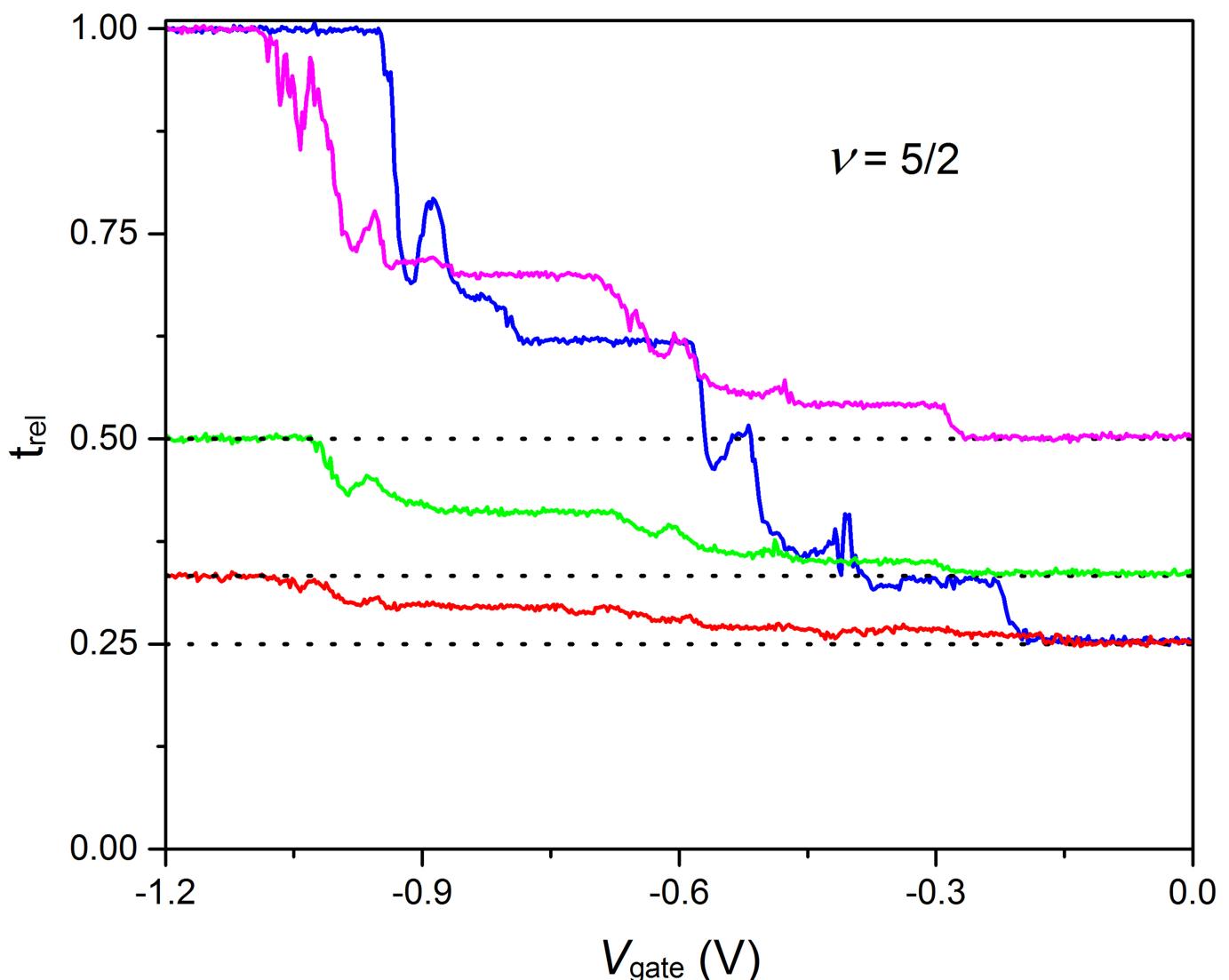
to the AlAs wells have low mobility and thus do not participate effectively in the conduction process. This structure suffers from substantial added bulk heat conductance. The structure used in our study, with δ -Si doping in low-Al-mole-fraction AlGaAs, did not have a visible bulk thermal conductance.



Extended Data Fig. 2 | Longitudinal resistance of the high-mobility SPSL-grown heterostructure. Longitudinal resistance measured in a Hall bar 100 μm wide and 200 μm long. Fractional filling factors are more

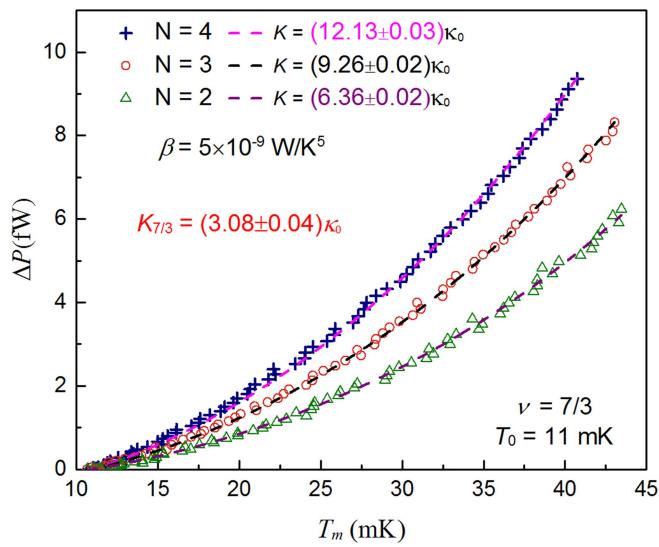
pronounced than in the δ -Si-doped structure. Yet, the structure suffers from added thermal conductance in the bulk (see main text).



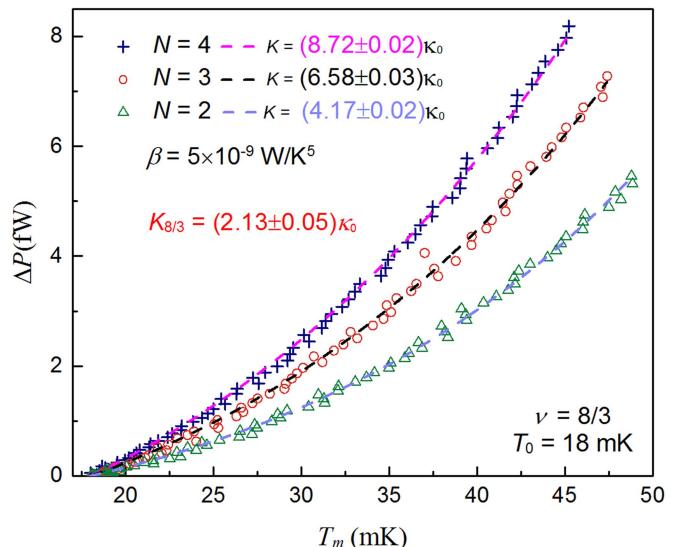


Extended Data Fig. 4 | Equal branching of current in all arms at $\nu = 5/2$. Current is sourced from the source, S, and measured in the drain, D, in the same arm (see Fig. 1a). The blue curve shows the reflection coefficient of the current measured in the drain as a function of the pinching of the arm gate. The reflection coefficient value starts from 0.25, when all the arm gates are fully open, and reaches 1.00, when all the current is reflected.

The red, green and magenta curves correspond to measurements for the fully open ‘measurement arm’, performed while the other arm gates deplete gradually one by one. Four open arms give a reflection coefficient of $r = 0.25$, whereas three open arms lead to $r = 0.33$ and two open arms give $r = 0.50$. The dotted lines are guides for the eyes indicating equal branching of currents.



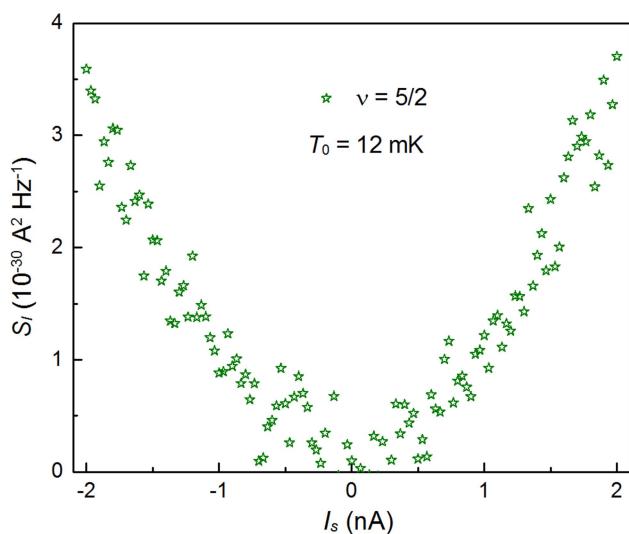
(a)



(b)

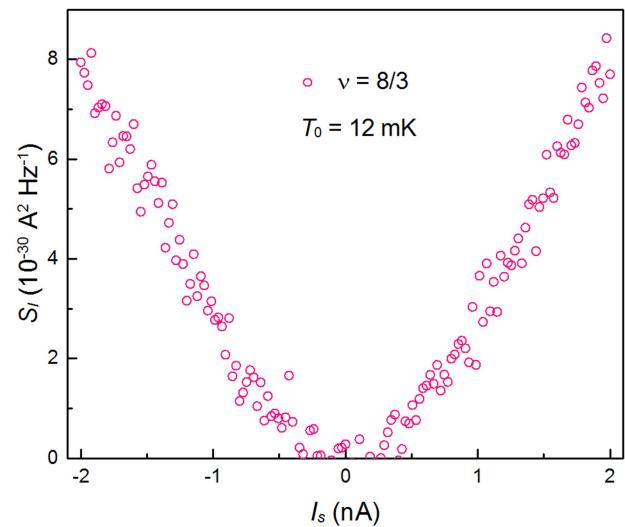
Extended Data Fig. 5 | Thermal noise analysis at $\nu = 7/3$ and $\nu = 8/3$.
a, b, Standard analysis (see main text), without subtracting the number of participating arms, but using the phonon contribution coefficient β , which fits extremely well in a large range of temperatures and at different filling

factors (errors of the fit are 99% confidence levels). The agreement with the expected data is clear. We note the added thermal heat conductance at $\nu = 8/3$ (b; see text).



(a)

Extended Data Fig. 6 | Upstream neutral modes in $\nu = 5/2$ and $\nu = 8/3$.
a, b, The noise measured at an upstream floating contact connected to



(b)

the cold amplifier (with respect to ground) is clear evidence of upstream neutral modes. Such upstream noise is not found in particle-like states²¹.

Kinase-controlled phase transition of membraneless organelles in mitosis

Arpan Kumar Rai¹, Jia-Xuan Chen^{2,4}, Matthias Selbach^{2,3} & Lucas Pelkmans^{1*}

Liquid–liquid phase separation has been shown to underlie the formation and disassembly of membraneless organelles in cells, but the cellular mechanisms that control this phenomenon are poorly understood. A prominent example of regulated and reversible segregation of liquid phases may occur during mitosis, when membraneless organelles disappear upon nuclear–envelope breakdown and reappear as mitosis is completed. Here we show that the dual-specificity kinase DYRK3 acts as a central dissolvase of several types of membraneless organelle during mitosis. DYRK3 kinase activity is essential to prevent the unmixing of the mitotic cytoplasm into aberrant liquid–like hybrid organelles and the over-nucleation of spindle bodies. Our work supports a mechanism in which the dilution of phase-separating proteins during nuclear–envelope breakdown and the DYRK3-dependent degree of their solubility combine to allow cells to dissolve and condense several membraneless organelles during mitosis.

A fundamental property of inhomogeneous crowded fluids is their ability to display coacervation, electrostatically driven liquid–liquid phase separation of molecules, which can lead to the emergence of micro-compartments^{1,2}. This phase-separation paradigm has in recent years been successfully applied to explain the formation of numerous membraneless organelles in cells, a process that is driven by intrinsically disordered regions in proteins and weak multivalent interactions^{3–5}, and often aided by polymeric scaffolds such as RNA to which these proteins bind^{6,7} (in some cases, the scaffolds themselves can also induce phase separation⁸). However, as with all cellular processes, various gene functions and enzyme activities must have evolved to modulate the underlying physico-chemical phenomenon, to drive it out of equilibrium or to push it across critical boundaries to make it advantageous for living systems.

We previously discovered that upon cellular stress, the dual-specificity kinase DYRK3 partitions into stress granules via its intrinsically disordered N-terminal domain and that its kinase activity is required for the dissolution of stress granules, probably by phosphorylating multiple RNA-binding proteins⁹. Also, it was shown that MBK-2, the *Caenorhabditis elegans* homologue of DYRK3, controls the dissolution of P-granules during the first cell division of the embryo¹⁰ through phosphorylation. In addition, POM1, a *Saccharomyces pombe* relative of DYRK3, displays reversible clustering¹¹ and controls the presence of protein assemblies involved in cytokinesis¹², which may also involve phase separation. Thus, DYRK-family kinases could represent a novel class of evolutionarily conserved cellular regulators that control phase-transition phenomena in cells, serving various cell-physiological purposes.

A notable example of such control may occur during mitosis. It has long been known that numerous membraneless organelles in both the cytoplasm and the nucleus disappear during mitosis, or, in the case of stress granules, cannot be induced to form in mitotic cells^{13–15}. However, the physical principles and the molecular mechanisms behind these phenomena have remained unclear. We here investigate whether DYRK3 acts as the dissolvase in these processes, and how this is coordinated with the cell cycle.

DYRK3 interactome and localization

To identify proteins that interact with DYRK3 in unperturbed cells, we performed quantitative affinity purification using stable-isotope labelling with amino acids in cell culture (SILAC)-based quantitative proteomics¹⁶. From two independent label-swap experiments (forward and reverse) (Extended Data Fig. 1a), we identified a total of 251 proteins that interact specifically with transiently overexpressed EGFP-tagged DYRK3 (Fig. 1a, Supplementary Table 1). The majority (86%) of these proteins are known to bind RNA, and amongst them are multiple components of stress granules, splicing speckles and the centrosome or pericentriolar matrix (Fig. 1a). Moreover, when we treated cells with GSK-626616, a small-compound inhibitor of DYRK3⁹, many of these interactions became more prominent (Fig. 1b, Extended Data Fig. 1b, c, Supplementary Table 2), independently of any increase in the abundance of the corresponding interactors or of DYRK3 itself (Extended Data Fig. 1d, e).

Overexpression of DYRK3 leads to its spontaneous phase separation at a specific concentration threshold⁹. To investigate its subcellular localization at near-endogenous levels, we created an inducible cell line. At very low induction levels (approximately 30% of endogenous levels) (Extended Data Fig. 2a), EGFP–DYRK3 shows a diffuse distribution within the nucleus and the cytoplasm, but also an enrichment in centrosomes (pericentrin) and the pericentriolar matrix (PCM1) (Fig. 1c), and in splicing speckles (SC35) (Fig. 1d). Inhibition of DYRK3 leads to a further accumulation in splicing speckles, and to a visible increase in the size of splicing speckles (Fig. 1d). Consistent with our previous findings⁹, EGFP–DYRK3 that is inhibited by GSK-626616 also accumulates in arsenite-induced stress granules (PABP1) at these low expression levels (Extended Data Fig. 2b). In mitotic cells, we observed that EGFP–DYRK3 localizes to spindle poles, and that it accumulates in small granules that grow in size when its kinase activity is inhibited (Fig. 1e, Extended Data Fig. 2c).

Mitotic effects of DYRK3 inhibition

Because DYRK3 associates with membraneless organelles in both the nucleus and the cytoplasm, accumulates in granules in mitotic cells when inhibited, and interacts with proteins that are heavily

¹Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. ²Max Delbrück Center for Molecular Medicine, Berlin, Germany. ³Charité-Universitätsmedizin Berlin, Berlin, Germany. ⁴Present address: Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. *e-mail: lucas.pelkmans@imls.uzh.ch

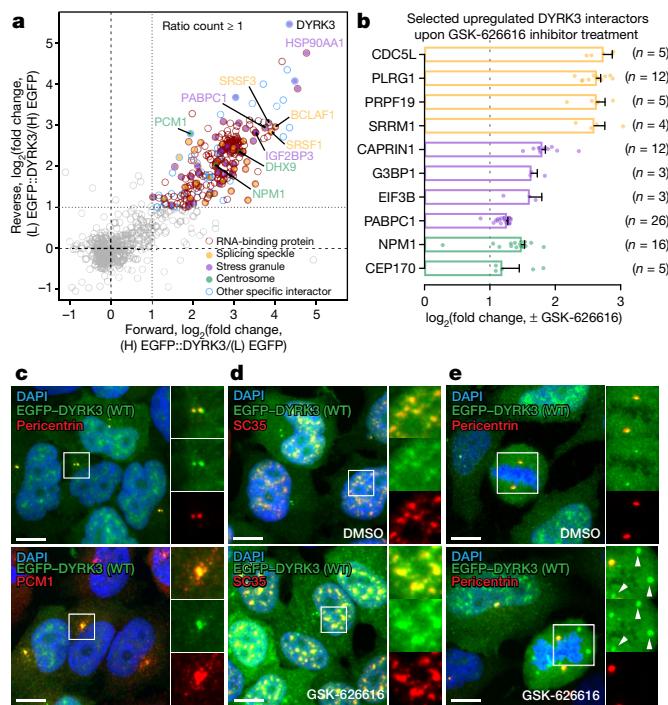


Fig. 1 | DYRK3 interacts with multiple membraneless compartments. **a**, Scatter plot of DYRK3-specific interactors determined by double-SILAC pull-down and mass spectrometry. H, SILAC heavy labelled; L, SILAC light labelled. **b**, Examples of DYRK3 interactions that increase upon GSK-626616 treatment. Error bars represent the standard error of the median. *n*, number of peptide evidence ratios included for calculating the protein ratio in one triple-SILAC pull-down. **c**, Colocalization of EGFP-tagged wild-type DYRK3 (EGFP-DYRK3(WT)) with centrosomes. **d**, Colocalization of EGFP-DYRK3(WT) with splicing speckles and the increase upon GSK-626616 treatment (1 μ M, 2 h). **e**, Formation of DYRK3-positive granules (arrowheads) upon GSK-626616 treatment (1 μ M, 4 h). Images are representative of at least three independent experiments. DAPI, 4',6-diamidino-2-phenylindole. Scale bars, 10 μ m.

phosphorylated during mitosis (some of which contain sites that are sensitive to DYRK3 inhibition⁹ (Extended Data Fig. 2d)), we decided to investigate its role during mitosis, when the contents of both cellular compartments mix. Notably, we found that inhibition of DYRK3 in mitotic cells leads to aberrant condensations of the splicing-speckle marker SC35 (Fig. 2a, Extended Data Fig. 3a), the stress-granule marker PABP (Fig. 2b, Extended Data Fig. 3b) and the pericentriolar-matrix protein PCM1 (Fig. 2c, Extended Data Fig. 3c). Aberrant condensations were observed in two different human cell lines and for multiple markers (Extended Data Fig. 3d-f). Testing a panel of kinase inhibitors that are known to regulate centrosome biogenesis or splicing-speckle composition in interphase cells did not show these effects (Extended Data Fig. 3g). Moreover, GSK-626616 does not cause the accumulation of SRPK1¹⁷ or CDK1, a key regulator of entry into mitosis, in these aberrant condensations, nor does it affect their activation or activity inside cells (Extended Data Fig. 4a-f) or in vitro⁹, whereas RNA interference-mediated knockdown of DYRK3 recapitulated the effects of GSK-626616 (Extended Data Fig. 4g).

Notably, splicing-speckle, stress-granule and pericentriolar-matrix markers co-condensed into the same hybrid structures, which also stained positive for polyadenylated RNA and accumulated inhibited DYRK3 (Fig. 2d, e, Extended Data Fig. 5a, b). Not all markers of membraneless organelles condensed into these structures, as P-bodies, nucleoli, and Cajal bodies appeared to dissolve normally during mitosis even when DYRK3 was inhibited (Extended Data Fig. 5c-e). The aberrant condensates are not classical aggresome-like structures, as they do not stain positive for ubiquitin (Extended Data Fig. 5f). Time-lapse analysis of mitotic granule formation using mCherry-tagged SRRM2,

a splicing-speckle protein, revealed that they appear fast (within 15–20 min) upon DYRK3 inhibition (Fig. 2f, Supplementary Video 1) and display liquid-like merging (Extended Data Fig. 6a), whereas photo-bleaching the granules showed that SRRM2 exchanges rapidly (with a recovery half-time of approximately 10 s) and completely between the condensed and dissolved phases (Fig. 2g). The total amount of SRRM2 in cells did not change during granule formation (Extended Data Fig. 6b), and the accumulation of SRRM2 in the condensed phase was a consequence of its depletion from the dissolved phase (Extended Data Fig. 6c). Thus, DYRK3 kinase activity is essential during mitosis to prevent the formation of aberrant liquid–liquid phase-separated hybrid condensates consisting of nuclear and cytoplasmic proteins and RNA, by keeping the condensation threshold of its substrates high.

DYRK3 dissolvase activity

To prevent aberrant condensation, DYRK3 may act as a dissolvase of multiple liquid-unmixed compartments. When we overexpressed wild-type EGFP-DYRK3 in interphase cells, we observed dissolution of splicing speckles in the nucleus (Fig. 3a, b, Extended Data Fig. 7a). This dissolving effect was dependent on its kinase activity, as it was reversed by treating cells with the DYRK3 inhibitor, and was not observed when overexpressing a kinase-dead point mutant of EGFP-DYRK3 (Fig. 3a, b, Extended Data Fig. 7a, b). Moreover, a nuclear-localization sequence mutant of DYRK3 prevented the dissolution of splicing speckles, whereas exclusive nuclear localization of DYRK3 completely dissolved splicing speckles, indicating that this sequence is required for the direct interaction with its substrates (Extended Data Fig. 7c, d).

Similarly, we found that overexpression of EGFP-DYRK3 dissolves pericentriolar satellites and prevents the induction of stress granules with arsenite in the cytoplasm in a kinase-activity-dependent manner (Fig. 3c-f, Extended Data 7e-h). Furthermore, phase-separated structures formed by overexpressing an intrinsically disordered domain of PCM1 (1–146)¹⁴ and SRRM1, a highly intrinsically disordered interactor of DYRK3 and known splicing-speckle component (Extended Data Fig. 7i), are completely dissolved by DYRK3 in a kinase-dependent manner (Extended Data Fig. 7j-l). We did not observe a dissolving effect of overexpressed DYRK3 on nucleoli (Extended Data Fig. 7m), consistent with the observation that these organelles still dissolve in DYRK3-inhibited mitotic cells. This shows that kinase-active DYRK3 has dissolvase activities for multiple but not all membraneless organelles, both in the cytoplasm and in the nucleus.

DYRK3-to-substrate ratio drives phase transition

Upon entry into mitosis, membraneless organelles disappear, and then reappear during telophase or after completion of mitosis^{13–15,18}. If disappearance occurs through a dissolution process driven by DYRK3, then its dissolvase activity must rapidly increase at the beginning of mitosis and then reduce at the end of mitosis. When we analysed the protein levels of endogenous DYRK3 along the cell cycle, we observed that its expression increases as cells progress from late S to the end of G2 (Fig. 4a). Yet, cells in late G2 display splicing speckles in the nucleus and can condense stress granules in the cytoplasm upon stress^{13,15,18}. Furthermore, we did not observe a sudden increase in DYRK3 levels as cells enter mitosis (black box in Fig. 4a).

One explanation could be that the concentration of DYRK3 relative to its substrate suddenly increases upon breakdown of the nuclear envelope. This occurs because key substrates of DYRK3 are exclusively located in either the nucleus (SRRM1) or the cytoplasm (PCM1), resulting in their dilution when the compartment barrier is removed (Fig. 4b, Extended Data Fig. 8a-d). And because DYRK3 itself is located in both compartments and does not get similarly diluted, the DYRK3-to-substrate ratio increases by a factor of two to three (Fig. 4b) as cells transition from G2 into M. To investigate whether such an increase in ratio is necessary and sufficient to ensure complete dissolution of membraneless organelles during mitosis, we transiently transfected inducible wild-type EGFP-DYRK3 along with mCherry-SRRM1 in non-mitotic cells. From a large number of single cells covering a wide

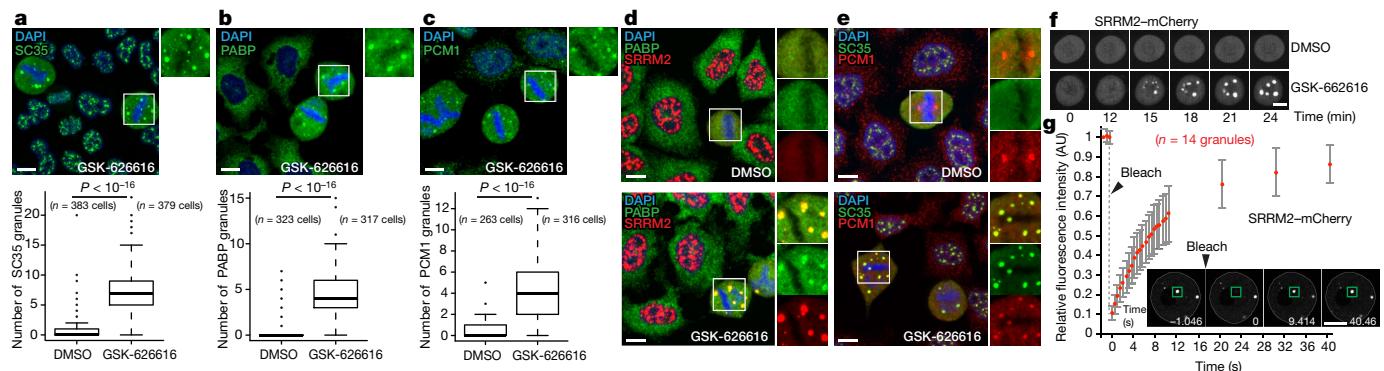
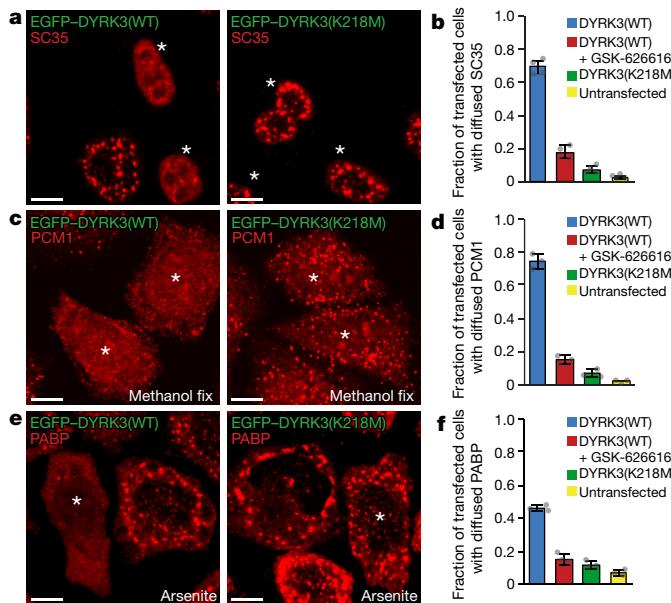


Fig. 2 | Formation of mitotic hybrid compartments upon DYRK3 inhibition. **a**, Top, mitotic SC35 granules upon GSK-626616 treatment (1 μ M, 1 h). Bottom, quantification of metaphase SC35 granule number. **b**, Top, mitotic PABP granules upon GSK-626616 treatment (1 μ M, 3 h). Bottom, quantification of metaphase PABP granule number. **c**, Top, mitotic PCM1 granules upon GSK-626616 treatment (1 μ M, 6 h). Bottom, quantification of metaphase PCM1 granule number. **d**, Colocalization of splicing and stress-granule markers in mitotic cells upon GSK-626616 treatment (1 μ M, 6 h). **e**, Colocalization of splicing and pericentriolar-satellite markers in mitotic cells upon GSK-626616 treatment (1 μ M, 6 h).

range of expression levels and DYRK3-to-SRRM1 ratios, we determined an intracellular phase diagram of SRRM1 in a condensed or dissolved phase as a function of DYRK3 and SRRM1 levels. This revealed a remarkably sharp phase boundary at a specific DYRK3-to-SRRM1 ratio across a large range of expression levels, below which SRRM1 would be condensed and above which it would be dissolved (Fig. 4c). Furthermore, by following interphase cells using time-lapse imaging, we could observe how changes in the concentration of the components determine granule phase behaviour as mapped in the phase diagram.



f, Time-lapse images show SRRM2-mCherry granule formation upon GSK-626616 (1 μ M) treatment in a mitotic cell. **g**, Fluorescence recovery after photobleaching (FRAP) trajectories of mitotic SRRM2-mCherry granules in the presence of GSK-626616 (1 μ M). Images show FRAP recovery. Data are mean \pm s.d. **a-c, g**, Box plots: centre line, median; box, interquartile range; whiskers, 1.5 \times interquartile range; dots, outliers. Statistical analysis was performed across cells using a Welch's two-sided *t*-test. Data in **a-c, g** are from three independent experiments. Images are representative of at least three independent experiments. AU, arbitrary units. Scale bars, 10 μ m.

Cells in which the DYRK3-to-SRRM1 ratio increased over time displayed a sudden transitioning from a condensed SRRM1 phase to a dissolved phase when the phase boundary was crossed (cells 1 and 2 in Fig. 4d). Consistently, SRRM1 remained dissolved in cells that always displayed a high DYRK3-to-SRRM1 ratio and stayed above the phase boundary over time (cell 3 in Fig. 4d).

We next focused on single cells with low, near-endogenous levels of EGFP-DYRK3 and mCherry-SRRM1 that entered mitosis during the time-lapse recording. When we plotted single-cell time traces of DYRK3-to-SRRM1 ratios into the phase diagram, and marked when SRRM1 becomes dissolved, we observed that this occurs right at the moment of crossing the phase transition boundary (cells 1 and 2 in Fig. 4e), owing to the sudden, 2.7-fold increase in DYRK3-to-SRRM1 ratio during the G2-to-M transition (marked by arrowheads in Fig. 4e). Consistently, when we followed a cell with high levels of overexpressed SRRM1 and a consequently very low DYRK3-to-SRRM1 ratio far away from the phase transition boundary, the 2.7-fold increase in the ratio during the G2-to-M transition was not sufficient to cross the boundary, resulting in the persistence of SRRM1 granules in mitotic cells (cell 3 in Fig. 4e). We observed this effect invariably in single cells that overexpress SRRM1 beyond a critical concentration, at which point levels of endogenous DYRK3 were insufficient to trigger a phase transition upon mitotic entry (Fig. 5a). These cells show mitotic SRRM1 granules (Fig. 5a) that recruit endogenous splicing-speckle proteins (Extended Data Fig. 8e), which also undergo liquid-like merging, and show rapid exchange of SRRM1 between the condensed and dissolved phase, which is under the control of DYRK3 (Extended Data Fig. 8f, g). Furthermore, we observed that the dilution effect alone is not sufficient. When we arrested cells in mitosis after nuclear-envelope breakdown with a single thymidine-nocodazole block, and splicing speckles are dissolved, acute treatment with the DYRK3 inhibitor results in the appearance of aberrantly condensed structures (Extended Data Fig. 8h), indicating that DYRK3 kinase activity is required to maintain the dissolved phase during mitosis.

Aberrant granules delay mitotic progression

We next asked whether the aberrant co-condensation of proteins in mitotic cells has a consequence for mitotic progression. We found that when cells are treated with the DYRK3 inhibitor, they display a prolonged mitotic length, and often a mitotic arrest (Fig. 5a). Notably, we also found a prolonged mitotic length in cells that were unable to completely dissolve mitotic granules upon nuclear-envelope breakdown owing to very high levels of overexpressed SRRM1, even when DYRK3

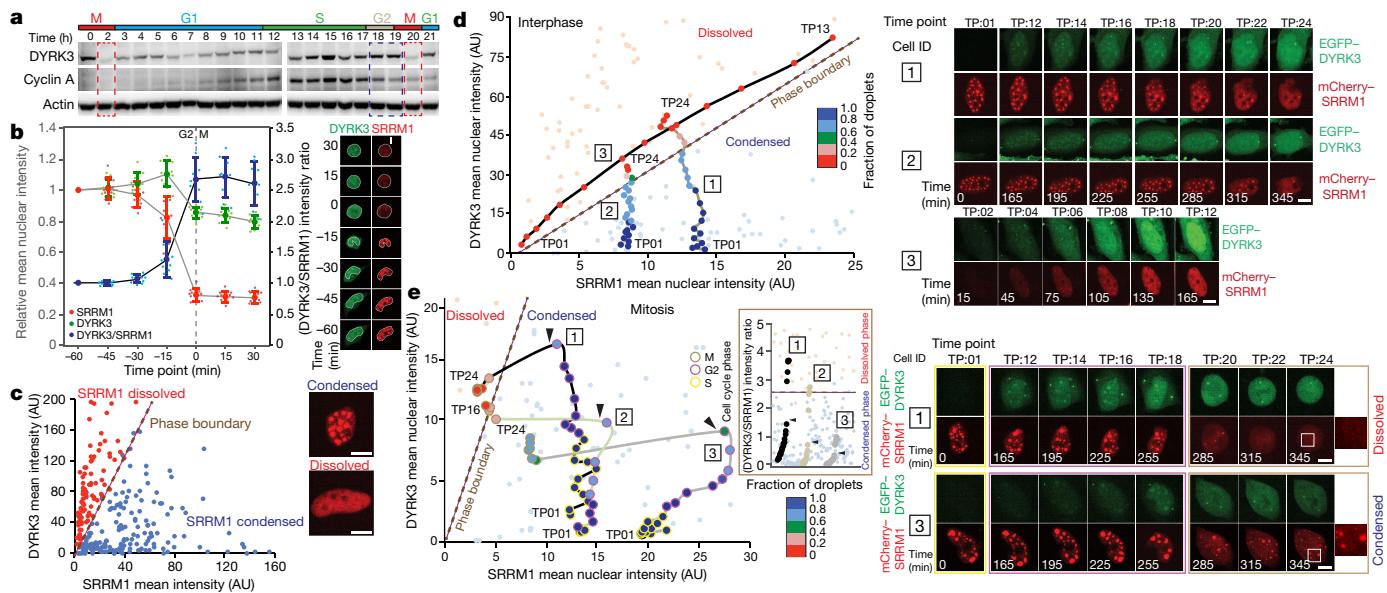


Fig. 4 | Increase in DYRK3-to-substrate ratio during mitosis drives dissolution of phase-separated compartments. a, Western blot analysis of endogenous DYRK3 levels in cells synchronized in mitosis and released for indicated times. **b**, Left, nuclear intensity change during the G2-to-M transition. EGFP-DYRK3(WT) (green), mCherry-SRRM1 (red) and EGFP-DYRK3(WT):mCherry-SRRM1 ratio (blue). Data are mean \pm s.d. from three independent experiments. The 0 min time point represents nuclear-envelope breakdown. Right, time-lapse images of a cell transitioning from G2 to M. **c**, EGFP-DYRK3(WT) and mCherry-SRRM1 nuclear intensities for a population of cells. Data points are colour-coded

activity was not compromised (Fig. 5a). This suggests that mitotic defects are, in part, caused by an inability to keep proteins dissolved in mitotic cells. Notably, although centrosome duplication appears to occur normally in DYRK3-inhibited cells, γ -tubulin displays multiple condensation foci (Extended Data Fig. 9a), resulting in the frequent occurrence of multipolar spindles (Fig. 5b). To explain this, we searched the DYRK3 interactome for possible mitotic regulators that might have an impact on mitosis. This revealed that ZNF207 (BuGZ) and its interaction partner BUB3 are enriched in the DYRK3 interactome upon GSK-626616 treatment (Fig. 5c, Supplementary Table 2). ZNF207 is required for spindle matrix assembly through liquid–liquid phase separation¹⁹, and localizes to splicing speckles in interphase cells²⁰ (Extended Data Fig. 9b). When we analysed its localization in mitotic cells, we found that upon DYRK3 inhibition, ZNF207 also becomes sequestered in mitotic granules in which it colocalizes with inhibited DYRK3 (Extended Data Fig. 9c, d) and PABP (Fig. 5c). Thus, DYRK3 releases mitotic regulators from liquid-unmixed compartments at the onset of mitosis and keeps them dissolved during mitosis. When this is compromised, mitotic regulators are sequestered in mitotic granules, which interferes with their normal roles during mitosis leading to mitotic defects.

At the end of mitosis, reassembly of the nuclear envelope and the reaccumulation of DYRK3 substrates in their respective compartments should cause the DYRK3 dissolvase-to-substrate ratio to drop below the critical point and reverse this process. However, some membraneless organelles, such as splicing speckles, recondense in telophase¹⁸, before the nuclear envelope has been reassembled. Notably, we noticed that DYRK3 is abruptly degraded at the boundary between M and G1 (red box in Fig. 4a and Fig. 5d), suggesting that recondensation before nuclear envelope reassembly may rely on rapidly removing DYRK3. Progression through mitosis and entry into G1 is driven by the anaphase promoting complex/cyclosome (APC/C), which promotes the ordered ubiquitination and subsequent degradation of multiple D-box (RXXL)-containing substrates through mitosis²¹. We found that DYRK3, which contains a D-box, interacts with both APC/C co-activators

for SRRM1 state. **d**, Left, trajectory of interphase cells expressing inducible EGFP-DYRK3(WT) and mCherry-SRRM1, mapped onto the phase diagram of Fig. 4c. Right, images of the cells plotted in the left panel; boxed numbers indicate matched cells. **e**, Left, trajectory of cells expressing inducible EGFP-DYRK3(WT) and mCherry-SRRM1 transitioning through cell-cycle stages, mapped onto the phase diagram of Fig. 4c. Inset, DYRK3:SRRM1 ratio plotted for the trajectories. Arrowheads indicate the G2-to-M transition. Right, images of the cells plotted in the left panel. Boxed numbers indicate matched cells. Data in Fig. 4a, c–e are representative of two independent experiments. Scale bars, 10 μ m.

CDC20 and CDH1, resulting in its targeting to ubiquitin-positive aggregates in the cytoplasm and its degradation (Fig. 5e and Extended Data Fig. 10a–d), consistent with the identification of DYRK3 as a novel APC/C substrate²². Moreover, when we overexpressed EGFP-DYRK3 in mitotic cells to levels that cannot be degraded by APC/C, splicing speckles remained dissolved in telophase, indicating that APC/C-mediated degradation of DYRK3 at the end of mitosis is involved in their recondensation process (Extended Data Fig. 10e).

Discussion

Here, we have uncovered a mechanism by which multiple membraneless organelles dissolve during mitosis and recondense as mitosis completes. The dual-specificity kinase DYRK3 has a key role in this process by acting as the dissolvase of these organelles. This extends our previous finding that DYRK3 dissolves stress granules during stress recovery⁹. Notably, not all membraneless organelles that disappear during mitosis, such as P-bodies and nucleoli, are dissolved by DYRK3. This may suggest a different physico-chemical phenomenon that underlies the formation of these organelles, or a role for other kinases, possibly relatives of DYRK3, that mediate their dissolution.

The organelles that are dissolved by DYRK3 all contain protein and RNA, and form by liquid–liquid phase separation. DYRK3 binds to proteins that are key components of these organelles, and is a proline-directed kinase with broad specificity²³ that phosphorylates multiple serine and threonine residues in unstructured domains⁹. This may affect the electrostatic properties of these domains such that it alters the condensation threshold of these proteins. When the active concentration of DYRK3 becomes high enough relative to its substrates, it triggers a transition from the condensed to the dissolved phase, which will exist as long as this ratio stays sufficiently high. The ratio can be modulated in multiple ways. At the onset of mitosis, the loss of the nucleus–cytoplasm compartment boundary is exploited to alter the DYRK3-to-substrate ratios such that it results in a phase transition. For some organelles, degradation of the dissolvase is used to trigger a phase transition in the opposite direction. Thus, progression through the cell

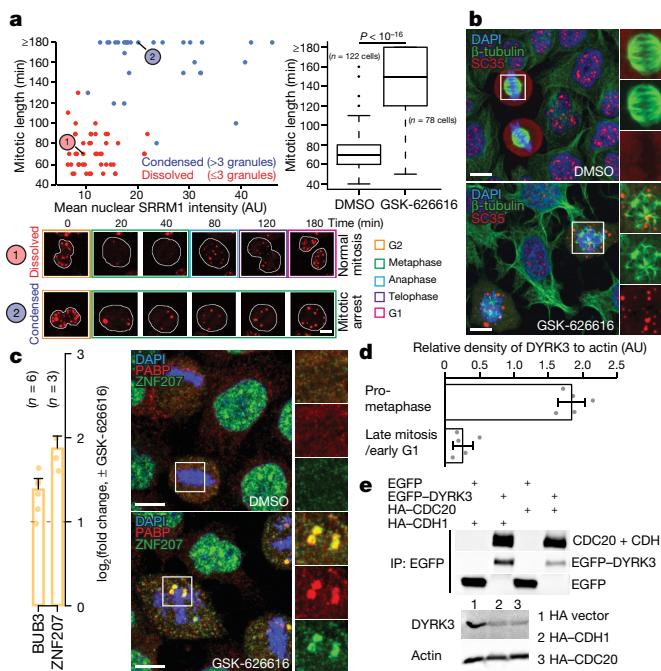


Fig. 5 | Mitotic defects upon inhibition of DYRK3 and APC/C-dependent degradation of DYRK3. **a**, Top left, mitotic length as a function of mCherry–SRRM1 nuclear intensity (in G2 phase). Data points are colour-coded for metaphase SRRM1 state. Top right, increase in mitotic length upon GSK-626616 treatment (1 μ M). Bottom, images of cells marked in the top left panel. **b**, Formation of multipolar spindles in mitotic cells upon GSK-626616 treatment (1 μ M, 6 h). **c**, Left, increase in the interaction of DYRK3 with BUB3 and ZNF207 upon GSK-626616 treatment. Error bars represent the standard error of the median. n indicates the number of peptide evidence ratios included for calculating the protein ratio in one triple-SILAC pull-down experiment. Right, colocalization of spindle-matrix and stress-granule markers in mitotic cells upon GSK-626616 treatment (1 μ M, 3 h). **d**, Western blot quantification of DYRK3 relative density during M and early G1 (M + 2 h). Data are mean \pm s.d. **e**, Top, DYRK3 interacts with the APC/C co-activators CDC20 and CDH1. Bottom, endogenous DYRK3 is degraded by overexpressed haemagglutinin-tagged CDH1 (HA–CDH1) and HA–CDC20. Box plots: centre line, median; box, interquartile range; whiskers, 1.5 \times interquartile range; dot, outliers. Statistical analysis performed across cells using a Welch's two-sided t -test. **a–c, e**, Data are from at least three independent experiments. **a–c, e**, Images are representative. Scale bars, 10 μ m.

cycle can be conceived as a process in which cells cycle through a phase diagram (Fig. 6). Also, when compartment boundaries stay intact, a specific accumulation or depletion of substrate in one compartment may alter the ratio sufficiently to induce a phase transition which may occur during cellular stress²⁴. In addition, the ratio could be altered by increasing the abundance or activity of DYRK3 by specific upstream regulators, which could be coupled to different physiological processes such as the cell cycle^{23,25}.

We currently do not understand how at the end of mitosis, and upon degradation of DYRK3, aberrant co-condensation is prevented. One possibility is that the kinetics or regulation of the dephosphorylation of DYRK3-phosphorylated residues in proteins is component- or compartment-specific, as may be the case for the levels of dephosphorylation required to initiate phase transition. In addition, cytoplasm and nucleus compartment boundaries are re-established at this time, which may contribute to preventing aberrant co-condensation.

Multiple aspects of mitosis are compromised when the DYRK3-to-substrate ratio cannot display its natural dynamics. Increases in the ratio allow the release of mitotic regulators from liquid–unmixed compartments as cells enter mitosis and prevent their aberrant co-condensation during mitosis. In addition, preventing co-condensation during mitosis could be important to ensure homogeneous distribution of proteins

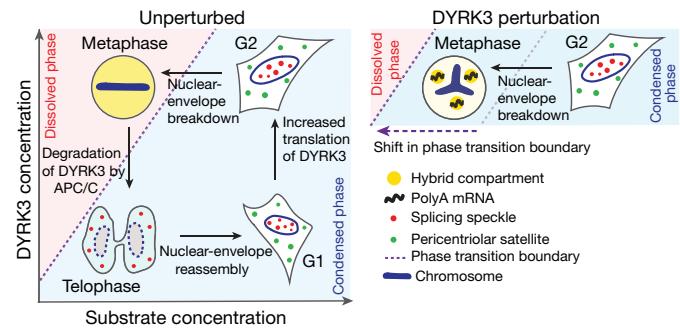


Fig. 6 | Model showing membraneless organelles transitioning through a phase diagram during the cell cycle. DYRK3 perturbation shifts the phase boundary resulting in the formation of hybrid compartments in mitotic cells.

and RNA between daughter cells during symmetric cell division, which may then be additionally regulated to achieve differential condensation points in daughter cells during asymmetric cell division^{3,10,26}. Finally, through its ability to control the condensation of multiple compartments, DYRK3 may link structural and functional roles for RNA, transcription and splicing to the assembly of the mitotic spindle^{27–29}, and be involved in the maintenance of the pericentriolar matrix. We propose that DYRK family kinases represent a novel, evolutionarily conserved class of regulators that controls liquid–liquid unmixing phenomena in cells, which are crucial for a variety of cell-physiological processes.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0279-8>

Received: 22 September 2017; Accepted: 15 May 2018;

Published online 4 July 2018.

- Overbeek, J. T. G. & Voorn, M. J. Phase separation in polyelectrolyte solutions. Theory of complex coacervation. *J. Cell. Comp. Physiol.* **49**, 7–26 (1957).
- Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
- Brangwynne, C. P. et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
- Kato, M. et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012).
- Li, P. et al. Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–340 (2012).
- Zhang, H. et al. RNA controls polyQ protein phase transitions. *Mol. Cell* **60**, 220–230 (2015).
- Saha, S. et al. Polar positioning of phase-separated liquid compartments in cells regulated by an mRNA competition mechanism. *Cell* **166**, 1572–1584.e16 (2016).
- Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
- Wippich, F. et al. Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling. *Cell* **152**, 791–805 (2013).
- Wang, J. T. et al. Regulation of RNA granule dynamics by phosphorylation of serine-rich, intrinsically disordered proteins in *C. elegans*. *eLife* **3**, e04591 (2014).
- Saunders, T. E. et al. Noise reduction in the intracellular pom1 gradient by a dynamic clustering mechanism. *Dev. Cell* **22**, 558–572 (2012).
- Rincon, S. A. et al. Pom1 regulates the assembly of Cdr2–Mid1 cortical nodes for robust spatial control of cytokinesis. *J. Cell Biol.* **206**, 61–77 (2014).
- Spector, D. L. & Smith, H. C. Redistribution of U-snRNPs during mitosis. *Exp. Cell Res.* **163**, 87–94 (1986).
- Dammermann, A. & Merdes, A. Assembly of centrosomal proteins and microtubule organization depends on PCM-1. *J. Cell Biol.* **159**, 255–266 (2002).
- Sivan, G., Kedersha, N. & Elroy-Stein, O. Ribosomal slowdown mediates translational arrest during cellular division. *Mol. Cell. Biol.* **27**, 6639–6646 (2007).
- Ong, S.-E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
- Gui, J. F., Lane, W. S. & Fu, X. D. A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature* **369**, 678–682 (1994).
- Thiry, M. Behavior of interchromatin granules during the cell cycle. *Eur. J. Cell Biol.* **68**, 14–24 (1995).
- Jiang, H. et al. Phase transition of spindle-associated protein regulate spindle apparatus assembly. *Cell* **163**, 108–122 (2015).

20. Wan, Y. et al. Splicing function of mitotic regulators links R-loop-mediated DNA damage to tumor cell killing. *J. Cell Biol.* **209**, 235–246 (2015).
21. Sivakumar, S. & Gorbsky, G. J. Spatiotemporal regulation of the anaphase-promoting complex in mitosis. *Nat. Rev. Mol. Cell Biol.* **16**, 82–94 (2015).
22. Merbl, Y. & Kirschner, M. W. Large-scale detection of ubiquitination substrates using cell extracts and protein microarrays. *Proc. Natl. Acad. Sci. USA* **106**, 2543–2548 (2009).
23. Aranda, S., Laguna, A. & de la Luna, S. DYRK family of protein kinases: evolutionary relationships, biochemical properties, and functional roles. *FASEB J.* **25**, 449–462 (2011).
24. Kedersha, N. L., Gupta, M., Li, W., Miller, I. & Anderson, P. RNA-binding proteins TIA-1 and TIAR link the phosphorylation of eIF-2 α to the assembly of mammalian stress granules. *J. Cell Biol.* **147**, 1431–1442 (1999).
25. Cheng, K. C. C., Klancer, R., Singson, A. & Seydoux, G. Regulation of MBK-2/DYRK by CDK-1 and the pseudophosphatases EGG-4 and EGG-5 during the oocyte-to-embryo transition. *Cell* **139**, 560–572 (2009).
26. Pellettieri, J., Reinke, V., Kim, S. K. & Seydoux, G. Coordinate activation of maternal protein degradation during the egg-to-embryo transition in *C. elegans*. *Dev. Cell* **5**, 451–462 (2003).
27. Blower, M. D., Feric, E., Weis, K. & Heald, R. Genome-wide analysis demonstrates conserved localization of messenger RNAs to mitotic microtubules. *J. Cell Biol.* **179**, 1365–1373 (2007).
28. Chan, F. L. et al. Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proc. Natl. Acad. Sci. USA* **109**, 1979–1984 (2012).
29. Grenfell, A. W., Heald, R. & Strzelecka, M. Mitotic noncoding RNA processing promotes kinetochore and spindle assembly in *Xenopus*. *J. Cell Biol.* **214**, 133–141 (2016).

Acknowledgements We thank members of the Pelkmans and Klemm laboratories for discussions. We further thank J. Michael Peters, A. Merdes and M. Polymenidou for reagents and acknowledge the assistance and support of the Center for Microscopy and Image Analysis, University of Zurich for performing FRAP experiments. EMBO and HFSP LTF supported A.K.R. J.-X.C was supported by an MDC-NYU PhD exchange program fellowship. L.P. is supported by the Swiss National Science Foundation and the University of Zurich.

Reviewer information *Nature* thanks A. Gladfelter, A. Hyman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.P. conceived the project. L.P. and A.K.R. wrote the paper. A.K.R. performed and analysed the data. J.X.C. performed SILAC pull-down experiments and bioinformatic data analysis and was supervised by M.S.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0279-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0279-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to L.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cell culture. HeLa cells were a gift from M. Zerial, HeLa-FlpIn-Trex³⁰ cells were a gift from I. Dikic and HEK293T cells were from ATCC (Molsheim Cedex). HeLa and HEK293T cells were maintained in 10-cm dishes in DMEM supplemented with 10% fetal bovine serum (Sigma-Aldrich) and L-glutamine (Sigma-Aldrich). HeLa-FlpIn-Trex cells were maintained in 10-cm dishes in DMEM supplemented with 10% tetracycline-free fetal bovine serum (Clonotech Laboratories), L-glutamine and blasticidin (1 μ g ml⁻¹, Santa Cruz). HeLa-FlpIn-Trex cells expressing EGFP-DYRK3(WT) were induced using doxycycline (Sigma) with indicated time and concentration. All cells were maintained in a humidified incubator at 37 °C under 5% CO₂. All cell lines were regularly tested for mycoplasma contaminations. Test results were negative. For imaging experiments, cells were grown in 96-well plates (Greiner Bio-One International).

Inhibitors. GSK-626616 (PubChem CID: 15981157) was obtained from Tocris Bioscience. DMSO was from Sigma-Aldrich (D2438). The following inhibitors were used in this study: GW 843682X (Tocris Bioscience, 2977), ZM 447439 (Tocris Bioscience, 2458), barasertib (Selleck Chemicals, S1147), volasertib (Selleck Chemicals, S2235), VX-680 (Selleck Chemicals, S1048), TCA 2317 HCl (Tocris Bioscience, 4066), KHC819 (Tocris Bioscience, 4262), KHC819 (Calbiochem, Merck Millipore, 219511), TG003 (Tocris Bioscience, 4336), SRPIN340 (Santa Cruz, sc-394310) and SRPIN340 (Tocris Bioscience, 5063). Prior to inhibitor treatments, cells were washed and serum-deprived for 2 h.

Generation of HeLa-FlpIn-Trex expressing EGFP-DYRK3(WT). HeLa-FlpIn-Trex cells were co-transfected with plasmids, pcDNA5/FRT/TO-EGFP-DYRK3(WT) and Flp recombinase expression vector (pOG44) for 48 h, and then selected with hygromycin (250 μ g ml⁻¹, Thermo Fisher Scientific) for two weeks.

Monoclonal cells were isolated by limiting-dilution technique, and then expanded. **SILAC cell culture.** HEK293T cells (ATCC) were cultivated using SILAC DMEM media (Life Technologies) containing 'light' Lys-0 (L-lysine-¹²C₆¹⁴N₂), Arg-0 (L-arginine-¹²C₆¹⁴N₄); 'medium' Lys-4 (L-lysine-¹²C₆¹⁴N₂D₄), Arg-6 (L-arginine-¹³C₆¹⁴N₄); or 'heavy' Lys-8 (L-lysine-¹³C₆¹⁵N₂), Arg-10 (L-arginine-¹³C₆¹⁵N₄) and supplemented with 10% dialysed fetal bovine serum (PAA Laboratories), 4 mM Glutamax and 1 mM sodium pyruvate. Cells were maintained in a humidified incubator under 5% CO₂ at 37 °C.

SILAC quantitative pull-down assays. To study interaction partners of DYRK3, a label-swap SILAC pull-down experiment was set up as depicted in Extended Data Fig. 1a. Plasmids containing EGFP only or EGFP-DYRK3(WT) were transiently transfected into SILAC-labelled HEK293T cells using the linear polyethylenimine (PEI) transfection reagent (Polysciences). Cells of different SILAC states were collected 48 h after transfection and lysed separately using a Dounce homogenizer in lysis buffer containing 25 mM Tris/HCl pH 7.4, 125 mM KCl, 1 mM MgCl₂, 1 mM EGTA/KOH pH 8.0, 5% glycerol, 1% Triton X-100 and freshly added protease inhibitor cocktail (Roche). The cleared lysates were then incubated with anti-EGFP agarose beads (Chromotek) at 4 °C for 90 min. After that, the beads were washed two times in buffer I (25 mM Tris/HCl pH 7.4, 125 mM KCl, 1 mM MgCl₂, 1 mM EGTA/KOH pH 8.0, 5% glycerol, 0.1% Triton X-100) with one wash in buffer II (1 mM Tris/HCl pH 7.4, 150 mM KCl, 1 mM MgCl₂). Beads of different SILAC states were combined during the final washing step. Proteins were then eluted by heating the beads to 90 °C in 8 M guanidinium chloride and precipitated in ethanol at 4 °C.

A separate triple-SILAC pull-down experiment was carried out to study the effect of the DYRK3 inhibitor (GSK-626616) on DYRK3 interactions as shown in Extended Data Fig. 1b. In brief, 24 h post-transfection, cells were further treated with 10 μ M GSK-626616 inhibitor for 2 h or not treated. The remaining steps of the pull-down experiment were then conducted as described above.

Precipitated proteins were resolubilized in 6 M urea/2 M thiourea buffer (10 mM HEPES pH 8.0). Proteins were then reduced by dithiothreitol and alkylated by iodoacetamide in the dark, with sequential digestion in solution using lysyl endopeptidase (Lys-C, Wako) for 3 h and trypsin (Promega) overnight at room temperature as reported previously³¹. Peptides were desalted and purified by solid phase extraction in C18 StageTips³².

Liquid chromatography with tandem mass spectrometry. Peptide separation was done online by reversed phase chromatography using an in-house packed column (inner diameter: 75 μ m; ReproSil-Pur C18-AQ 3- μ m resin, Dr. Maisch GmbH) through a 120-min gradient of acetonitrile (8–50%) with 0.1% formic acid at a nanoflow rate of 200 nl min⁻¹. Eluted peptides were sprayed directly by electrospray ionization into a Q Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific). Mass spectra were acquired in data-dependent mode using a top 10 sensitive method. Each duty cycle consisted of one full scan (resolution: 70,000, target value: 3×10^6 , scan range: 300 to 1,700 m/z) and ten tandem mass spectrometry scans of fragment ions produced via higher energy collision dissociation (HCD; resolution: 35,000, target value: 5×10^5 , maximum injection time: 120 ms, isolation window: 4.0 m/z). Precursor ions with unassigned or +1 charge state were not selected for fragmentation scans. Dynamic exclusion time was 30 s.

Mass spectrometry data processing. All mass spectrometry raw data were processed using MaxQuant software package (v.1.5.5.1)³³ with the built-in Andromeda search engine³⁴. Spectral data were searched against a target-decoy database containing the forward and reverse protein sequences of UniProt human proteome release 2016_06 (92,578 entries), EGFP and the default list of 245 common contaminants. Corresponding SILAC labels were selected for the double- or triple-SILAC experiment. A maximum of three SILAC-labelled amino acids were allowed for each peptide. Trypsin/P specificity was assigned. Carbamidomethylation of cysteine was chosen as fixed modification. Methionine oxidation and protein N-terminal acetylation were set as variable modifications. A maximum of two missed cleavages were tolerated. Minimum peptide length of seven amino acids was required. Each protein group must contain at least one unique peptide. False discovery rate (FDR) was set to 1% for both peptide and protein identifications.

For SILAC protein quantification, minimum ratio count was set to one. Both the unique and razor peptides were used for quantification. The 're-quantify' function was switched on. The 'advanced ratio estimation' option was also chosen.

Bioinformatic data analysis. All data analyses were performed in the R statistical environment (v.3.3.1). To normalize the pull-down data, protein ratios were first log-transformed. Under the assumption that the majority of detected proteins were non-specific background binders, the distribution of protein ratios was assessed using kernel density estimation to find the peak density. Protein ratios were then normalized by adjusting to the point at which the peak density was found. For the triple-SILAC pull-down data, proteins were further filtered to retain only those with a normalized heavy labelled/light labelled or medium labelled/light labelled \log_2 ratio greater than the cut-off value of one. The log-transformed heavy labelled/medium labelled ratios were then re-scaled with respect to the DYRK3 heavy labelled/medium labelled ratio (Extended Data Fig. 1c).

Gene Ontology annotation for categorizing the detected DYRK3 interaction partners was downloaded on 28th June 2016 from the QuickGO database³⁵ provided by the European Bioinformatics Institute. The list of stress-granule components was obtained from previously published work³⁶.

To assess phosphorylation level changes of detected DYRK3 specific binders during the cell cycle, quantitative phosphoproteomic data of the human cell cycle were downloaded from a previous study³⁷. The data were filtered sequentially to retain: (1) phospho-sites with high confidence in localization probability (class I)³⁸; (2) phospho-sites that were regulated over cell cycle (changed by at least two folds from highest to lowest amounts); and (3) phospho-sites belonging to proteins that were identified as DYRK3-specific interaction partners in our study. The phospho-site ratios already normalized by changes in protein abundance were then re-scaled by z-score transformation and plotted as a heat map (Extended Data Fig. 2d).

The standard error of the median in Fig. 1 and Fig. 5 was computed on the basis of the bootstrap (1,000 times) sampling distribution of corresponding log-transformed ratios of peptide evidence. The protein ratio of ZNF207 in Fig. 5c is determined by a regression analysis using the 'advanced ratio estimation' function in MaxQuant.

Prediction of the protein low complexity region (LCR) was performed using the SEG algorithm with default parameters³⁹.

Cell synchronization. Cells were synchronized in mitosis (M) by treatment with 2 mM thymidine (Sigma-Aldrich) for 22 h, washed three times with PBS, released from thymidine block for 5 h in fresh medium and then treated with nocodazole (25 μ g ml⁻¹) (Sigma-Aldrich) for 14 h. For time-course experiments, mitotic cells were harvested by shake-off, and then centrifuged for 3 min at 300g. Cells were then washed three times with normal medium and released in normal medium for the indicated times.

Plasmids. Generation of full-length DYRK3 plasmids, pcDNA5-EGFP-DYRK3(WT) and pcDNA5-EGFP-DYRK3(K218M) has been reported previously⁹. pcDNA5-EGFP-NLS(SV40)-DYRK3(WT) and pcDNA5-EGFP-NLS(SV40)-DYRK3(K218M) constructs were generated by annealing oligonucleotides 5'-CCGGTTACCGAAG AAGAAGCGAAAGGTACA-3' and 5'-TTACCGAAGAAGAACGAAAGG TACACCGG-3', and ligating the product into pcDNA5-EGFP-DYRK3(WT) and pcDNA5-EGFP-DYRK3(K218M) plasmids using AgeI restriction site. The NLS(SV40) is the nuclear localizing signal derived from simian virus (SV40) large T antigen. The pcDNA5-EGFP-NLSmut-DYRK3 plasmid was generated by site-directed mutagenesis in pcDNA5-EGFP-DYRK3(WT) using primer pair 5'-GCTTGTGGGGTCGCTCAGCTGGGGTGCAGCGGGGGTC CCCCAGGCAGC-3' and 5'-GTCGCTGGGGACCCGCCGCTGCACCC GCAGCTGAGCGACCCCAACAGC-3'. mCherry-SRRM1 was generated by amplifying SRRM1 from Flag-SRRM1⁴⁰ (Addgene plasmid, 11305) using primer pair 5'-GCTAACGCTTCGATGGACGCGGGATTTTCCGGGAAC-3' and 5'-CATGGTACCTTATTAAGACTGTGGGGACACTTGGGCCTTC-3', and inserting the product into the mCherry2-C1 plasmid (gift from M. Davidson, Addgene plasmid, 54563) using HindIII and KpnI restriction sites. The pcDNA5/FRT/TO-EGFP-DYRK3(WT) vector was generated

by amplifying EGFP-DYRK3 from pcDNA5-EGFP-DYRK3(WT) using primer pair 5'-GGTGGTACCCGCCACCATGGTGAGCAAGGGCG-3' and 5'-GCTGGATCCTTATTAGCTAATCAGTTGGCAATACACT-3', and inserting the product into the pcDNA5/FRT/TO plasmid⁹ using KpnI and BamHI restriction sites. RFP-DYRK3(WT) and RFP-DYRK3(K218M) were generated by amplifying them from pcDNA5-EGFP-DYRK3(WT) and pcDNA5-EGFP-DYRK3(K218M), respectively, using primer pair 5'-ATCGAATTCTATGAAGTGGAAAGAGAACTGGGGATG-3' and 5'-GGTACCTTATTAGCTAATCAGTTGGCAATACAC-3', and inserting the products into the mRFP-C1 plasmid (a gift from R. Campbell, M. Davidson and R. Tsien, Addgene plasmid 54764) using EcoRI and KpnI restriction sites. Full-length SRRM2 (NM_016333.3) was synthesized by Genscript and inserted into the mCherry2-N1 plasmid (gift from M. Davidson, Addgene plasmid 54517) using KpnI and AgeI restriction sites. EGFP-PCM1 (1-1468) was a gift from A. Merdes. EGFP-3XNLS¹¹ was a gift from D. Mullins (Addgene plasmid, 58468). pmScarlet_NES_C1¹² was a gift from D. Gadella (Addgene plasmid, 85060). 1436 pcDNA3-Flag-HA was a gift from W. Sellers (Addgene plasmid, 10792). HA-CDH1⁴³ and HA-CDC20⁴³ were a gift from M. Kirschner (Addgene plasmid, 11596).

Plasmid transfections. Transfections were performed with the indicated plasmids using either GeneJuice reagent (EMD Millipore) or Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's instructions.

RNA fluorescent in situ hybridization (FISH). Fluorescent in situ hybridization (FISH) was performed with ATTO 488-labelled 18-nucleotide long oligo-dT probe (Microsynth) for detecting polyA-mRNA. Probe hybridization was performed using the reagents and protocol provided by the manufacturer (LGC Biosearch Technologies). Cells were grown in 96-well plates, fixed by adding 4% PFA for 10 min, and permeabilized with 70% ethanol for 6 h at 4 °C. Cells were then washed with wash buffer, and incubated with the probe in hybridization buffer for 6 h at 37 °C. The cells were washed again with wash buffer and then processed for immunofluorescence.

RNA interference. Cells were reverse-transfected using Lipofectamine RNAiMax (Thermo Fisher Scientific) according to manufacturer instructions, and grown in 96-well plates. Three silencer-select short interfering RNA (siRNA) against DYRK3 (5'-CGGAUUUUGGAGCAUCUUA-3', 5'-CCAUCUAGCUUAUCGAUUA-3' and 5'-GAAAGACAUUUGGAGGUUAU-3') were pooled together (Thermo Fisher Scientific). Positive (KIF11) and negative control silencer-select siRNA were from Thermo Fisher Scientific. 1 pmol siRNA was added per well of a 96-well plate. Cells were fixed 72 h post transfection.

Imaging. Imaging was carried out on an automated spinning disk microscope from Yokogawa (Cell Voyager 7000), with an enhanced CSU-X1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type), a 40× 0.95 NA Olympus objective, a 60× 1.2 NA Olympus water-immersion objective and Andor sCMOS cameras (Andor, 2,560× 2,160 pixels). Live-cell imaging experiments were performed with Yokogawa (Cell Voyager 7000) with a 60× 1.2 NA Olympus water-immersion objective. All cells were maintained in a humidified environment at 37 °C under 5% CO₂ for live-cell imaging experiments. For nucleo-cytoplasmic dilution experiments (Fig. 4b, Extended Data Fig. 8a-d), HeLa-FlpIn-Trex cells were transfected with the indicated plasmids for 6 h, synchronized in G1/S using thymidine (2 mM, 22 h), and then released into normal medium for 6 h before imaging. EGFP-DYRK3 expression was induced with doxycycline (1 µg ml⁻¹) for 2 h before imaging. Hoechst 33342 (0.5 µg ml⁻¹, Thermo Fisher Scientific) was added to stain DNA, 1 h before imaging. Six Z-stacks (1 µm apart) were acquired at each time point and data was stored as individual stacks. For other experiments, images were stored as maximum intensity projection. For mitotic arrest experiments, cells were transfected with indicated plasmids overnight, followed by thymidine treatment (2 mM, 22 h), and then released into normal medium for 6 h. Cells were then treated with nocodazole (25 ng ml⁻¹) for 6 h in serum-deprived medium to arrest cells in pro-metaphase. For experiments measuring the mean intensity of SRRM2-mCherry during mitotic granule formation (Extended Data Fig. 6b, c), cells were imaged with 60× 1.2 NA objective, and acquired Z stacks (0.45 µm apart, 43 stacks) were stored as individual files. For SRRM2-mCherry and mCherry-SRRM1 granule fusion experiments in mitotic cells, acquired Z stacks (1 µm apart, 11 stacks) were stored as individual files. For experiments to calculate the mitotic length, Hoechst 33342 (0.5 µg ml⁻¹, Thermo Fisher Scientific) was added to stain DNA, 1 h before imaging. Time-lapse imaging was performed with a time interval of 10 min.

Image analysis. Images were processed using ImageJ (<https://rsb.info.nih.gov/ij/>). For quantifying mitotic granules (SC35 granules, PABP granules and PCM1 granules) in metaphase, cells were manually segmented and mean cell intensity was determined for SC35, PABP and PCM1. Granules were determined using the FIJI 3D Objects Counter plugin, with an intensity threshold of 2.5× the mean mitotic cell intensity. We implemented an adjoining pixel cut-off of (8 pixels for SC35 granules, 12 pixels for PABP granules and 16 pixels for PCM1 granules; pixel size 161.3 nm × 161.3 nm)⁸. The 90th percentile of granule numbers for

the DMSO control data in Fig. 2a, b were used as the cut-off for quantification of Extended Data Fig. 3g (three for SC-35 granules and one for PABP granules). Mitotic granules were observed and quantified in metaphase cells (identified on the basis of DAPI staining).

For nuclear dilution experiments of EGFP-3×NLS, mCherry-SRRM1 and EGFP-DYRK3, the nuclei of cells in G2 phase were segmented on the basis of Hoechst signal. For cytoplasmic dilution experiments of pmScarlet-NES and EGFP-DYRK3, nuclear segmentation was performed on the basis of Hoescht signal, and the cell segmentation was performed using the cytoplasmic fluorescent signal from these proteins. For cytoplasmic dilution experiments of PCM1 (1-1468), cells were co-transfected with a plasmid expressing mCherry protein. Nuclear segmentation was performed on the basis of Hoescht signal, and the cell segmentation was performed using the cytoplasmic fluorescence signal of mCherry. Mitotic cells were segmented using the cytosolic fluorescence signal of the respective proteins. Mean nuclear or cytoplasmic intensity at each time-point is the mean intensity from four Z-stacks (1 µm apart). The mean nuclear intensity at each time point was normalized to the intensity at the -60-min time-point. For experiments measuring the mean intensity of SRRM2-mCherry during mitotic granule formation (Extended Data Fig. 6b, c), sum projection of the Z-stacks (0.45 µm apart, 43 stacks) was performed in ImageJ. Mitotic cells that were entirely captured within the Z-stacks during the course of the time-lapse movie were used for quantification.

For experiments performed in Extended Data Fig. 1d, cell and nuclear segmentation was performed using CellProfiler⁴⁴. Mitotic, border, missegmented cells, and cells with multiple nuclei were discarded using the supervised machine-learning tool, CellClassifier⁴⁵. Mean intensity features were extracted from segmented cells and used for quantification.

Mitotic length was calculated from prophase/pro-metaphase to anaphase on the basis of Hoescht signal. All data analyses were performed in Microsoft Excel and the R statistical environment. Exposure and image processing has been performed similarly for all sub-figures in each panel. Images have been rescaled similarly in all sub-figures in a panel except Fig. 4e. In Fig. 4e the images of mCherry-SRRM1 in M phase are rescaled differently compared to S and G2 phase for better visualization of condensed and dissolved states of SRRM1. Background subtraction was performed on images by calculating the mean background intensity outside the cells.

Statistical analysis was performed using Welch's two-sided *t*-test in the R statistical environment using the default *t*-test function.

Phase diagram experiment. For the phase diagram experiment (Fig. 4c-e), HeLa-FlpIn-Trex cells were co-transfected with pcDNA5/FRT/TO-EGFP-DYRK3(WT) and mCherry-SRRM1 for 24 h and then induced with doxycycline (1 µg ml⁻¹, Thermo Fisher Scientific) was added to stain DNA, an hour before imaging. Time-lapse imaging was performed with a time interval of 15 min.

The phase diagram (Fig. 4c) was built by randomly selecting cells from multiple time points from the start, middle and end of the time-lapse movie. Nuclear segmentation was manually performed on the basis of Hoescht signal, and the mean nuclear intensity of EGFP-DYRK3 and mCherry-SRRM1 was calculated from the maximum intensity projections. The nuclear intensities were used to determine the phase diagram. Cells were classified on the basis of SRRM1 granules being in a dissolved (≤3 granules) or condensed state (>3 granules). Cells in interphase (Fig. 4d) or cells transitioning from interphase to mitosis (Fig. 4e) from the same time-lapse experiment were mapped on the determined phase diagram.

FRAP analysis. HeLa-FlpIn-Trex cells were grown on 8-well chambers (Ibidi GmbH, Germany). Cells were transfected with the indicated plasmids for 24 h, for experiments in interphase cells. Photobleaching experiments in interphase cells were performed in serum-deprived medium. For mitotic experiments, cells were transfected overnight with indicated plasmids, with G1-S arrest using thymidine (2 mM). mCherry-SRRM1 transfected cells were released in serum-deprived medium and monitored during mitosis. SRRM2-mCherry transfected cells were arrested in mitosis using nocodazole (in serum-deprived medium). Photobleaching was performed on a Leica SP5 Mid UV-VIS equipped with a 63× 1.4 NA, oil, Plan-Apochromat objective. Cells were maintained at 37 °C under 5% CO₂ during the course of experiment. A defined region was bleached twice at full laser power. Recovery could not be monitored for long durations because of mitotic granules moving out in the Z-direction. Recovery curves are therefore plotted for durations in which the granules seemed not to move in the Z-direction. FRAP analysis was performed using ImageJ.

Immunofluorescence. Cells were grown in 96-well plates and fixed by adding 4% PFA (Electron Microscopy Sciences) for 20 min and then permeabilized with 0.2% Triton X-100 for 20 min. The cells were blocked with 1% BSA in PBS (blocking buffer), incubated with primary antibodies in blocking buffer for 2 h at room temperature, and then incubated with Alexa-Fluor labelled secondary antibody (Life Technologies) in blocking buffer for 1 h at room temperature. Nuclei were stained using DAPI (Life Technologies). Staining for γ-tubulin and

pT446 APC3 antibodies, was performed by pre-permeabilizing the cells with 0.1% Triton X-100 (in PBS) for 3 min before PFA fixation. Staining of ZNF207 (BugZ) was performed by washing cells with PEM buffer (100 mM PIPES, pH 6.8, 5 mM EGTA, 2 mM MgCl₂) and permeabilizing them with 0.05% Triton X-100 (in PEM buffer) for 3 min before PFA fixation⁴⁶ (in PEM buffer). Staining for PCM1 antibody (Cell Signaling, 5259S), was performed by fixing cells with −20°C methanol for 5 min and then blocking with 1% BSA. The following antibodies were used: PABP (1:100, Santa Cruz, sc-32318), G3BP (1:500, Abcam, ab56574), pericentrin (1:400, Abcam, ab4448), SC35 (1:400, Sigma-Aldrich, S4045), PCM1 (1:100, Santa Cruz, sc-67204), PCM1 (1:200, Cell Signaling, 5259S), SRRM2 (1:600, Sigma-Aldrich, HPA041411), γ-tubulin (1:200, Sigma-Aldrich, T6557), SON (1:600, Sigma-Aldrich, HPA023535), SRPK1 (1:200, BD Biosciences, 611072), cyclin B1 (1:800, Cell Signaling, 12231), CDK1 (1:200, Cell Signaling, 9116), phospho-CDK1 (Tyr15) (1:50, Cell Signaling, 4539), β-tubulin (1:600, Abcam, ab6046), DDX6 (1:400, Bethyl Laboratories, A300-461A), coilin (1:400, Abcam, ab87913), fibrillarin (1:400, Abcam, ab5821), ubiquitin (FK2) (1:200, Enzo Life Sciences, BML-PW8810), B-23 (1:600, Sigma-Aldrich, B0556), ZNF207 (1:500, Sigma-Aldrich, HPA017013), TIAR (1:200, BD Biosciences, 610352, provided by M. Polymenidou) and pT446 APC3 (1:200, provided by J. Michael Peters).

Western blotting. Samples were prepared for western blotting as follows: cells in culture dishes were washed three times with PBS before addition of lysis buffer (150 mM NaCl, 50 mM HEPES, 1% Triton X-100, 0.1% SDS, 2 mM DTT, 5 mM EDTA and 2× protease inhibitor cocktail in Milli-Q H₂O) and scraping. Lysates were incubated for 30 min on ice (with 15 s sonication every 15 min). The cells were then centrifuged at 21,000g for 15 min at 4°C and the supernatant was stored at −80°C. Proteins in the lysates were denatured by the addition of loading buffer and boiling for 10 min. Proteins were resolved by 10% SDS-PAGE and further analysed using immunoblotting. After wet transfer of proteins onto a PVDF membrane (Immobilon-P, 0.45 μm, Millipore), membranes were incubated in 5% low-fat milk in 1× PBST (1× PBS with 0.1% Tween-20) for 1 h at room temperature. Membranes were then probed with primary antibodies in 3% BSA in 1× PBST for 2 h at room temperature, and then incubated with HRP-conjugated secondary antibodies (1:5,000) in 5% low-fat milk in 1× PBST for 1 h at room temperature. For the detection of DYRK3, PABP and G3BP, membranes were incubated with primary antibody for 24 h at 4°C. Signal was revealed using HRP substrate solution. For the endogenous DYRK3 degradation experiment (Fig. 5e, bottom panel), 3 μg of HA, HA-CDC20 and HA-CDH1 plasmids were transfected for 48 h before the preparation of cell lysates. The following antibodies were used: DYRK3 (1:1,000, abcam, ab155949), EGFP (1:1,000, Cell Signaling, 2956), pan-actin (1:1,000, Cell Signaling, 8456), cyclin-A2 (1:500, Abcam, ab7956), PABP (1:500, Santa Cruz, sc-32318), G3BP (1:1,000, Abcam, ab56574), HA (1:1,000, Sigma Aldrich, 11867423001) and EGFP (1:1,000, Cell Signaling, 2956). For quantification, images were processed using ImageJ.

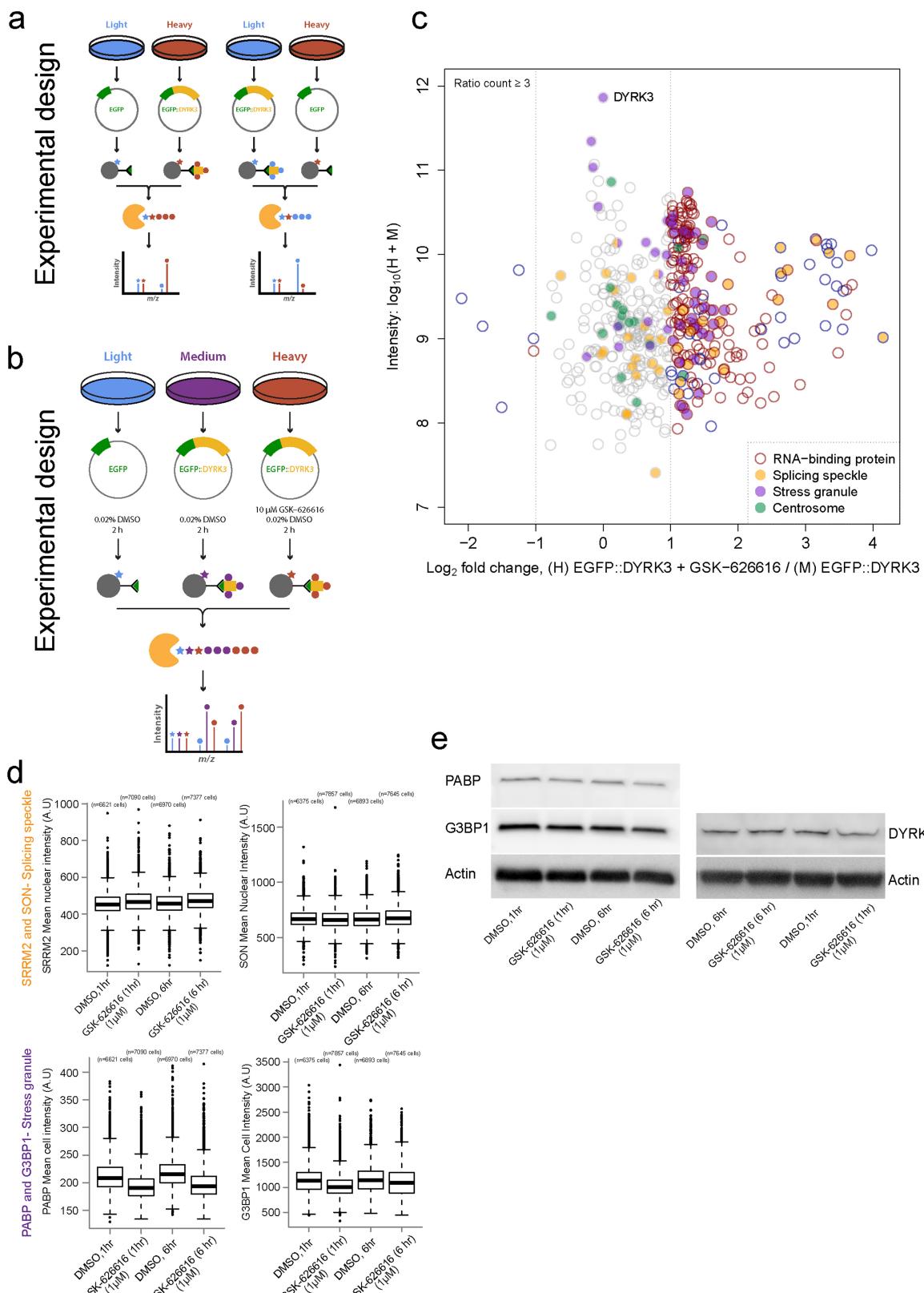
Immunoprecipitation experiments. Cells were collected in lysis buffer (150 mM NaCl, 50 mM HEPES, 1% Triton X-100, 0.1% SDS, 5 mM EDTA and 2× protease inhibitor cocktail in Milli-Q water), and allowed to lyse on ice for 20 min. The cells were then centrifuged at 21,000g for 15 min at 4°C and the supernatant was used for immunoprecipitation experiments. The cleared lysates were diluted tenfold in dilution buffer (150 mM NaCl, 50 mM HEPES and 5 mM EDTA and 1× protease inhibitor cocktail), and then incubated with anti-EGFP magnetic agarose beads (Chromotek) at 4°C for 120 min on a rotary shaker. The beads were then washed

three times with wash buffer (150 mM NaCl, 50 mM HEPES and 5 mM EDTA). To pull down the immunocomplexes, beads were boiled in 30 μl of 2× SDS-PAGE sample buffer for 15 min. The immunoprecipitated proteins were separated by SDS-PAGE. Western blot analysis was performed as described above. For Fig. 5e (top panel), HA-CDC20 and HA-CDC20 were probed with an anti-HA antibody. EGFP and EGFP-DYRK3 were probed with the anti-EGFP antibody.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

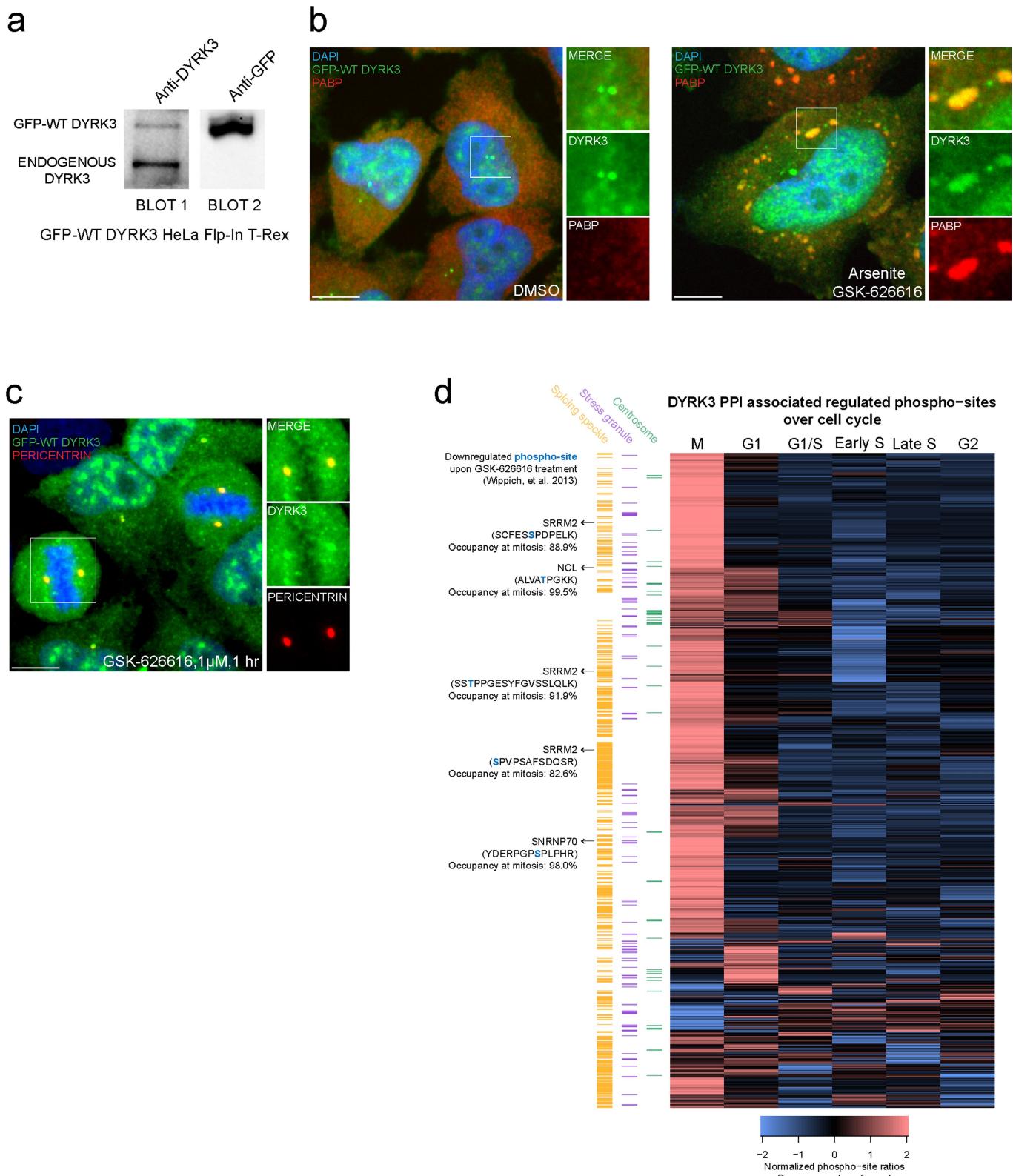
Data availability. The raw data that support the findings of this study are available from the corresponding author upon reasonable request. Mass spectrometry data are available via ProteomeXchange (<http://www.proteomexchange.org/>) with identifier PXD007761. Supplementary Information, which includes uncropped western blot images, and Source Data for Figs. 2, 3, 4, 5 and Extended Data Figs. 3, 4, 6, 8 are available with the online version of this paper.

30. Tighe, A., Staples, O. & Taylor, S. Mps1 kinase activity restrains anaphase during an unperturbed mitosis and targets Mad2 to kinetochores. *J. Cell Biol.* **181**, 893–901 (2008).
31. Paul, F. E., Hosp, F. & Selbach, M. Analyzing protein–protein interactions by quantitative mass spectrometry. *Methods* **54**, 387–395 (2011).
32. Rappaport, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).
33. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
34. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
35. Binns, D. et al. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009).
36. Aulas, A. & Vande Velde, C. Alterations in stress granule dynamics driven by TDP-43 and FUS: a link to pathological inclusions in ALS? *Front. Cell. Neurosci.* **9**, 423 (2015).
37. Olsen, J. V. et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* **3**, ra3 (2010).
38. Olsen, J. V. et al. Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648 (2006).
39. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).
40. Cheng, C. & Sharp, P. A. Regulation of CD44 alternative splicing by SRM160 and its potential role in tumor cell invasion. *Mol. Cell. Biol.* **26**, 362–370 (2006).
41. Belin, B. J., Cimini, B. A., Blackburn, E. H. & Mullins, R. D. Visualization of actin filaments and monomers in somatic cell nuclei. *Mol. Biol. Cell* **24**, 982–994 (2013).
42. Bindels, D. S. et al. mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nat. Methods* **14**, 53–56 (2017).
43. Pfleger, C. M., Lee, E. & Kirschner, M. W. Substrate recognition by the Cdc20 and Cdh1 components of the anaphase-promoting complex. *Genes Dev.* **15**, 2396–2407 (2001).
44. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
45. Rämö, P., Sacher, R., Snijder, B., Begemann, B. & Pelkmans, L. CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics* **25**, 3028–3030 (2009).
46. Jiang, H. et al. A microtubule-associated zinc finger protein, BuGZ, regulates mitotic chromosome alignment by ensuring Bub3 stability and kinetochore targeting. *Dev. Cell* **28**, 268–281 (2014).



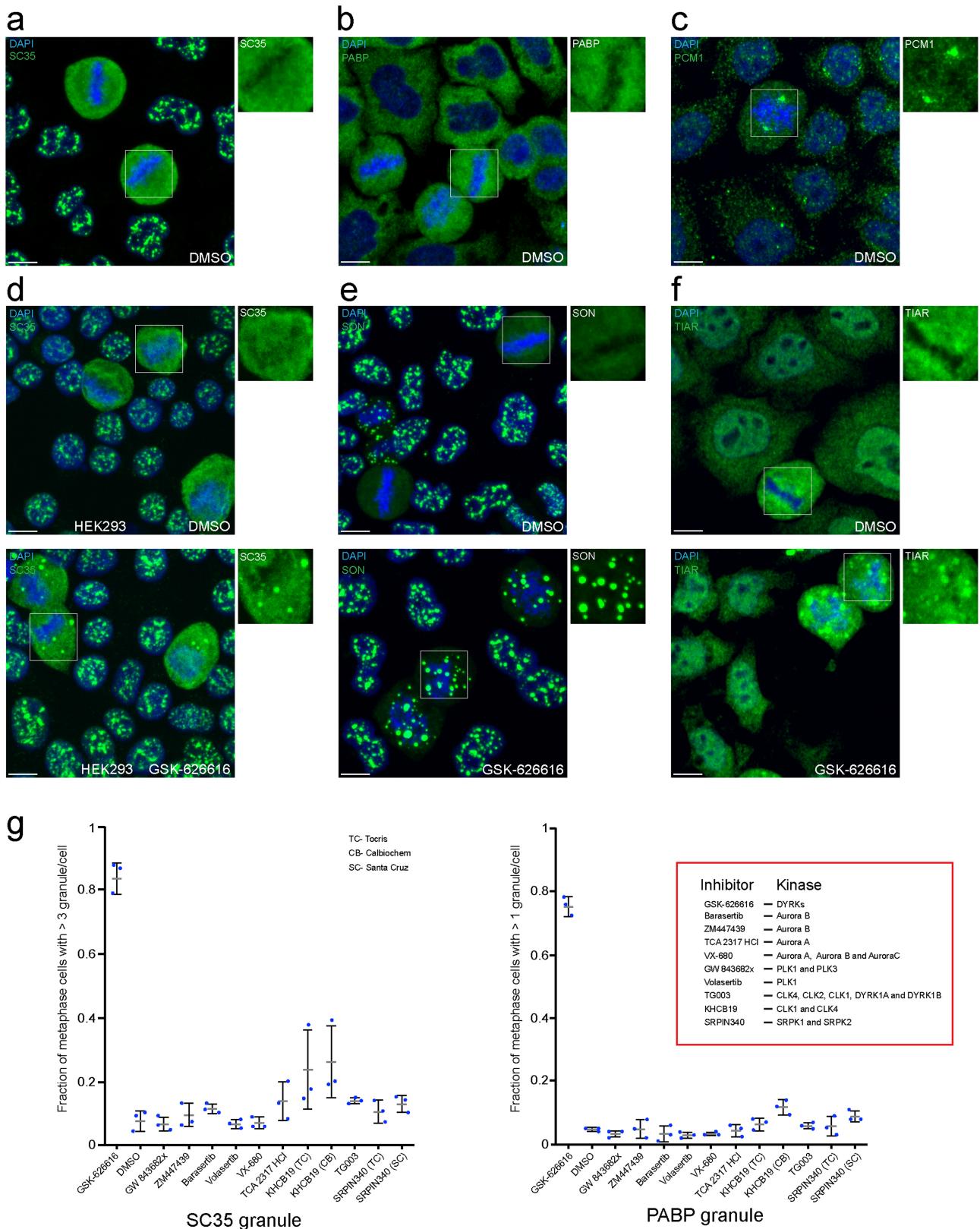
Extended Data Fig. 1 | DYRK3 interactions are differentially regulated upon GSK-626616 treatment. **a**, Experimental design of the SILAC quantitative pull-down assays to identify protein–protein interaction partners of DYRK3. **b**, Experimental design of triple-SILAC quantitative pull-down assays to identify change of DYRK3 interactors upon GSK-626616 inhibitor treatment. **c**, Scatter plot shows change in DYRK3 interactors upon GSK-626616 inhibitor treatment (normalized heavy labelled/medium labelled \log_2 ratios). Only proteins detected as DYRK3-specific interactors (normalized heavy labelled (H)/light labelled (L) or

medium labelled (M)/light labelled \log_2 ratio >1) were retained on this plot. Interactors are considered differentially regulated (coloured rims) if the heavy labelled/medium labelled \log_2 ratio is greater than 1 or less than -1 . **d**, Mean intensity of DYRK3 interactors upon GSK-626616 inhibitor treatment. Data are representative of two technical replicates. **e**, Protein abundance of DYRK3 interactors and DYRK3 upon GSK-626616 inhibitor treatment. Data are representative of two independent experiments. Box plots: centre line, median; box, interquartile range; whiskers, $1.5 \times$ interquartile range; dots, outliers.



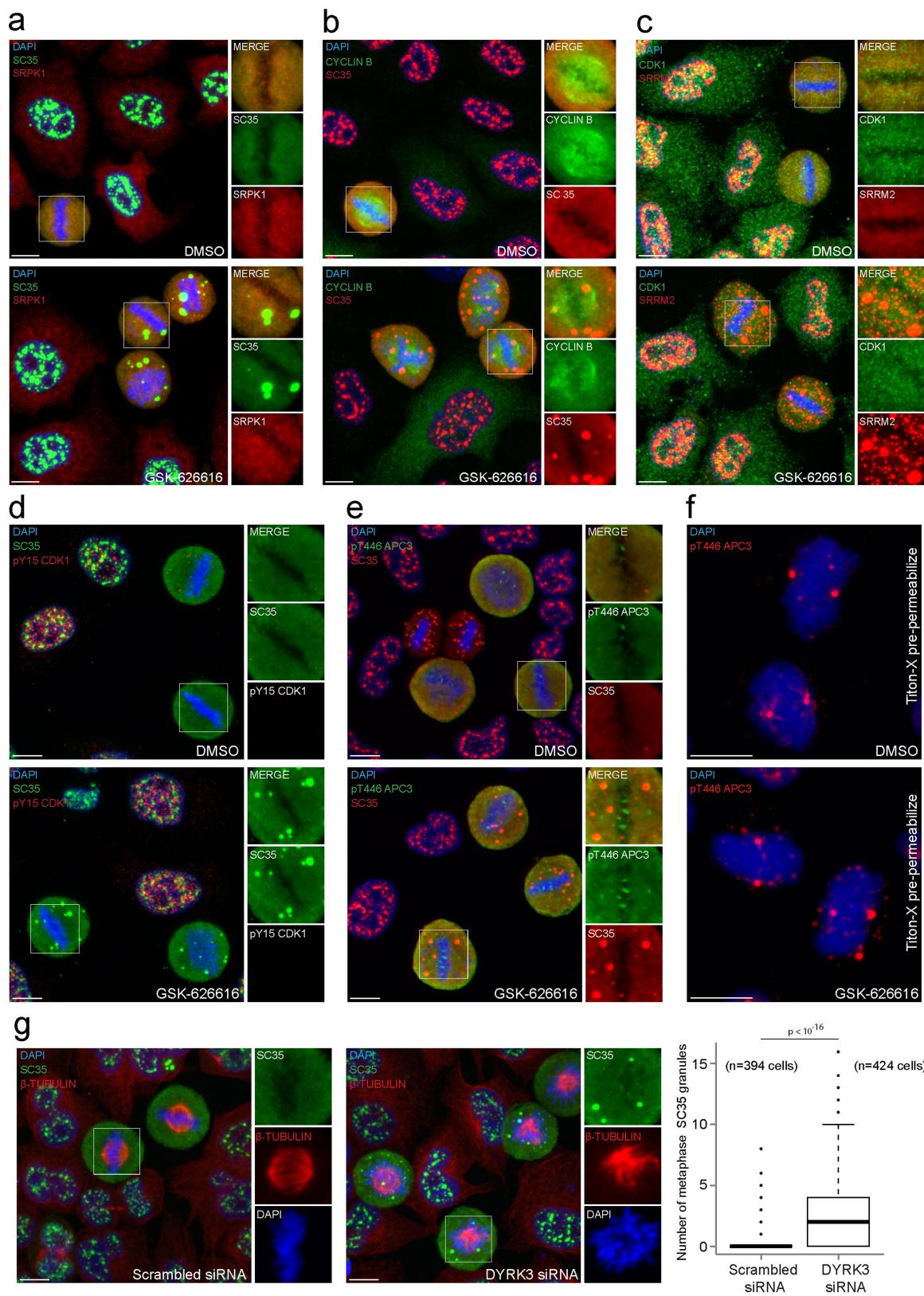
Extended Data Fig. 2 | GSK-626616-inhibited DYRK3 localizes to stress granules and forms mitotic granules. a, Western blot shows the expression level of endogenous DYRK3 and EGFP-DYRK3 in HeLa-FlpIn-Trex cells stably expressing inducible EGFP-DYRK3(WT). EGFP-DYRK3 expression was induced with doxycycline (500 ng ml⁻¹, 4 h). The same induction conditions were used for Fig. 1c–e and Extended Data 2b. **b**, Colocalization of GSK-626616-inhibited (1 μM, 2 h) EGFP-DYRK3(WT) with stress granules upon arsenite treatment (500 μM, 45 min). **c**, Mitotic cells show formation of small DYRK3-positive granules upon GSK-626616 treatment (1 μM, 1 h). **d**, Phosphoproteomic data of the

human cell cycle showing changes in the regulated phospho-sites that are associated with the DYRK3-specific interactors detected in Fig. 1a. The majority (74.4%) of these phospho-sites reached peak levels at mitosis. The phospho-site occupancy level is displayed for those sites that were downregulated upon GSK-626616 inhibitor treatment as previously reported⁹. The phosphoproteomic data were retrieved from a previously published work³⁷. The known localization (splicing speckle, stress granule or centrosome) of the corresponding protein is indicated for each phospho-site. Images and western blots are representative of at least three independent experiments. Scale bars, 10 μm.



Extended Data Fig. 3 | DYRK3 inhibition leads to the formation of aberrant mitotic granules. **a**, Dissolved staining of a splicing-speckle marker (SC35) in mitotic cells upon DMSO treatment. **b**, Dissolved staining of stress-granule marker (PABP) in mitotic cells upon DMSO treatment. **c**, Spindle-pole localization of pericentriolar material protein (PCM1) in mitotic cells upon DMSO treatment. **d**, Mitotic cells (HEK293T) show formation of SC35 granules upon GSK-626616 treatment (1 μ M, 1 h). **e**, Formation of SON granules in mitotic cells upon GSK-626616 treatment (1 μ M, 6 h). **f**, Formation of TIAR granules

in mitotic cells upon GSK-626616 treatment (1 μ M, 6 h). **g**, Left, quantification of fraction of mitotic (metaphase) cells with SC35 granules on treatment with kinase inhibitors (1 μ M, 1 h). Right, quantification of fraction of mitotic (metaphase) cells with PABP granules on treatment with kinase inhibitors (1 μ M, 3 h). Inhibitors and the corresponding kinase targets are mentioned. Data are mean \pm s.d. from three technical replicates. Images are representative of at least three independent experiments. Scale bars, 10 μ m.

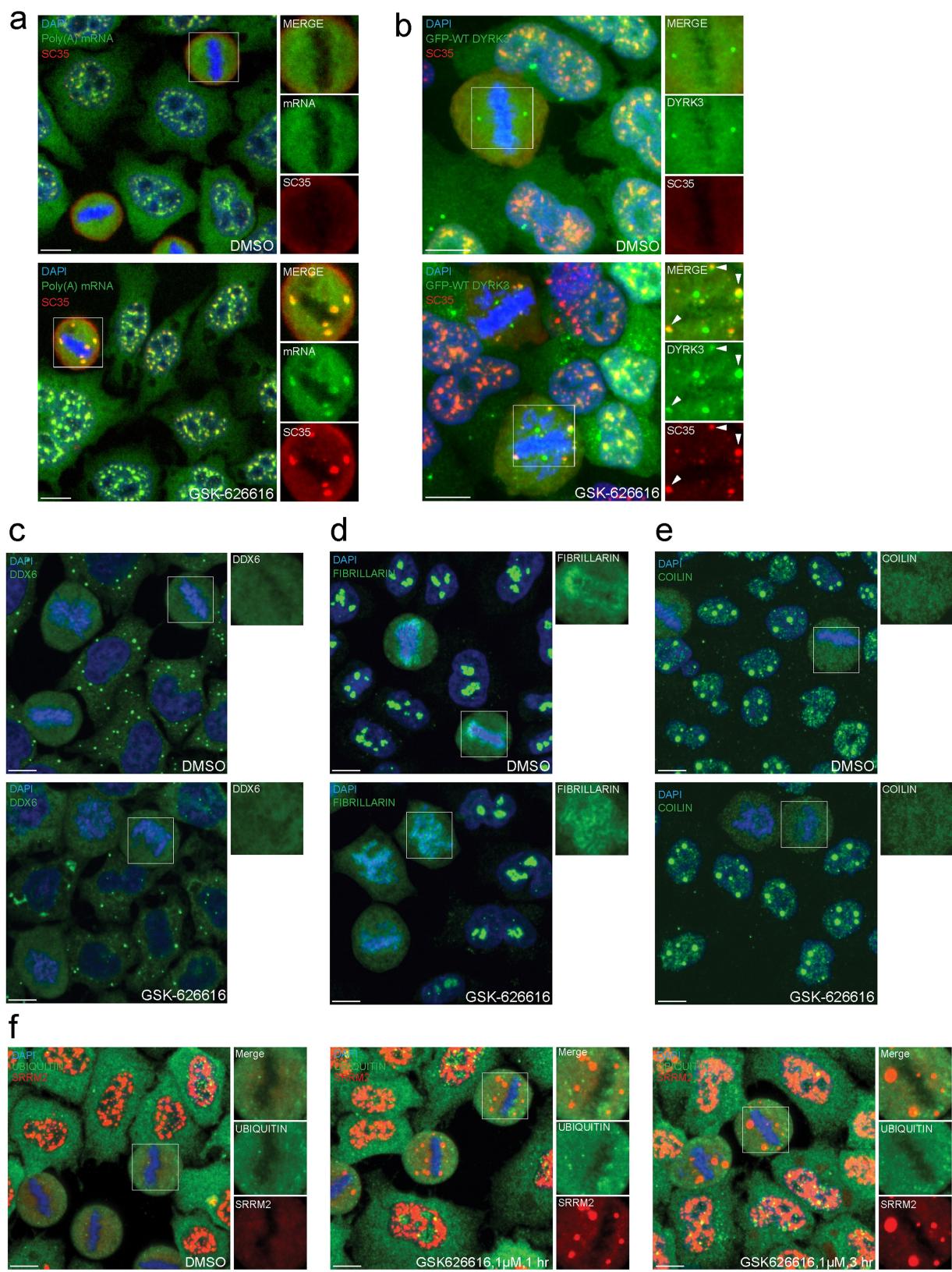


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Formation of mitotic granules is DYRK3 specific.

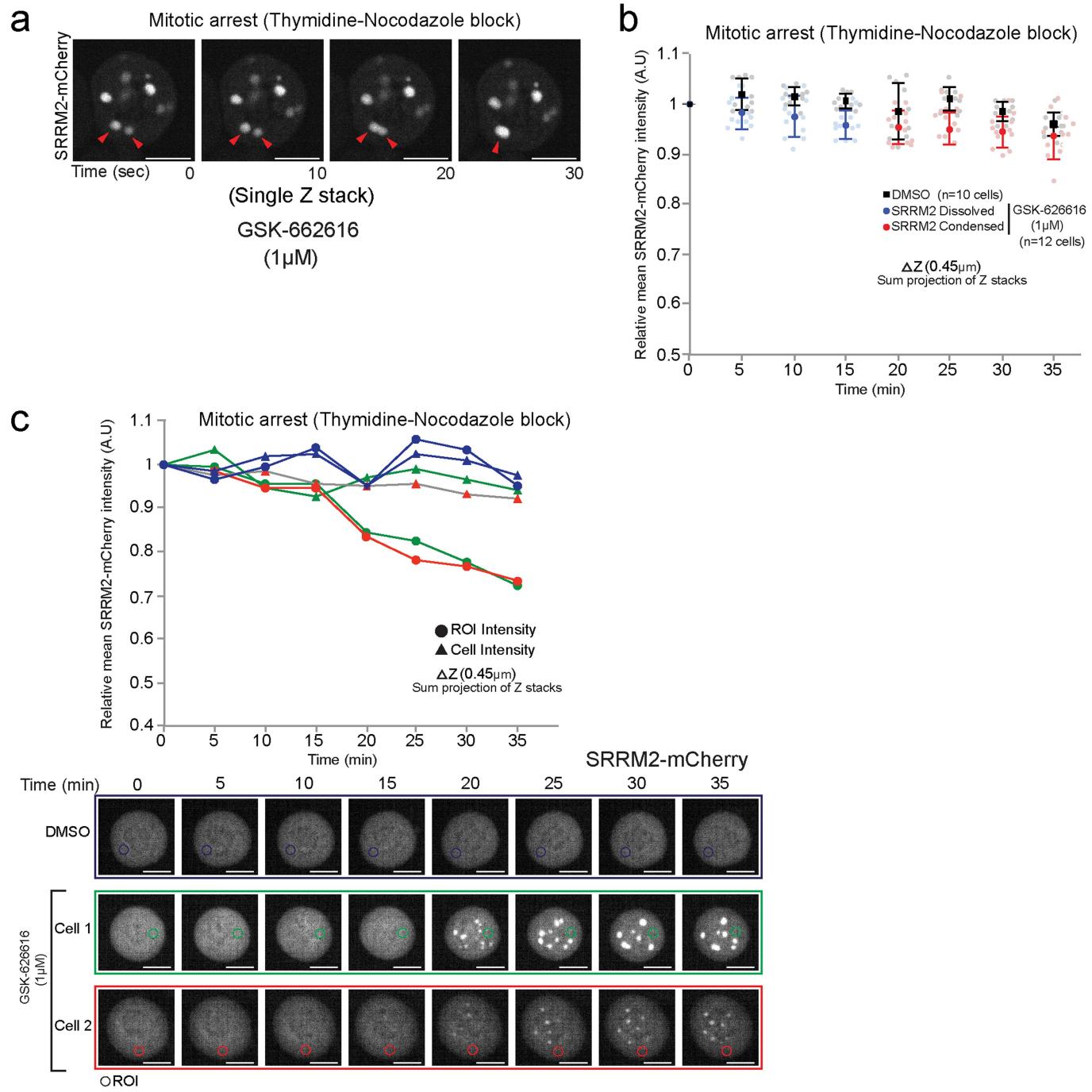
a, Mitotic cells show no colocalization between the splicing-speckle marker (SC35) and SRPK1 upon GSK-626616 treatment (1 μ M, 1 h). **b**, Mitotic cells show no colocalization between the splicing speckle marker (SC35) and cyclin B upon GSK-626616 treatment (1 μ M, 1 h). **c**, Mitotic cells show no colocalization between the splicing-speckle marker (SRRM2) and CDK1 upon GSK-626616 treatment (1 μ M, 1 h). **d**, Mitotic cells stained for pY15 CDK1. Loss of pY15 signal (CDK1 activation) in mitotic cells upon GSK-626616 (1 μ M, 1 h) treatment is comparable to DMSO control. **e**, GSK-626616 treatment (1 μ M, 1 h) does not result in a decrease in pT446 APC3 (CDK1 mitotic substrate) signal in mitotic cells compared to

the DMSO control. **f**, Mitotic cells show staining for pT446 APC3 (CDK1 mitotic substrate). The cells were pre-permeabilized with Triton-X before fixation. pT446 APC3 signal can be observed at spindle poles for both DMSO and GSK-626616 treatment (1 μ M, 1 h). **g**, Appearance of SC35 granules and spindle apparatus defects upon DYRK3 knockdown. Right, quantification of SC35 granule number in mitotic (metaphase) cells (four independent experiments). Box plots: centre line, population median; box, interquartile range; whiskers, 1.5 \times interquartile range; dots, outliers. Statistical analysis performed across cells using a Welch's two-sided *t*-test. Images and data are representative of at least three independent experiments. Scale bars, 10 μ m.



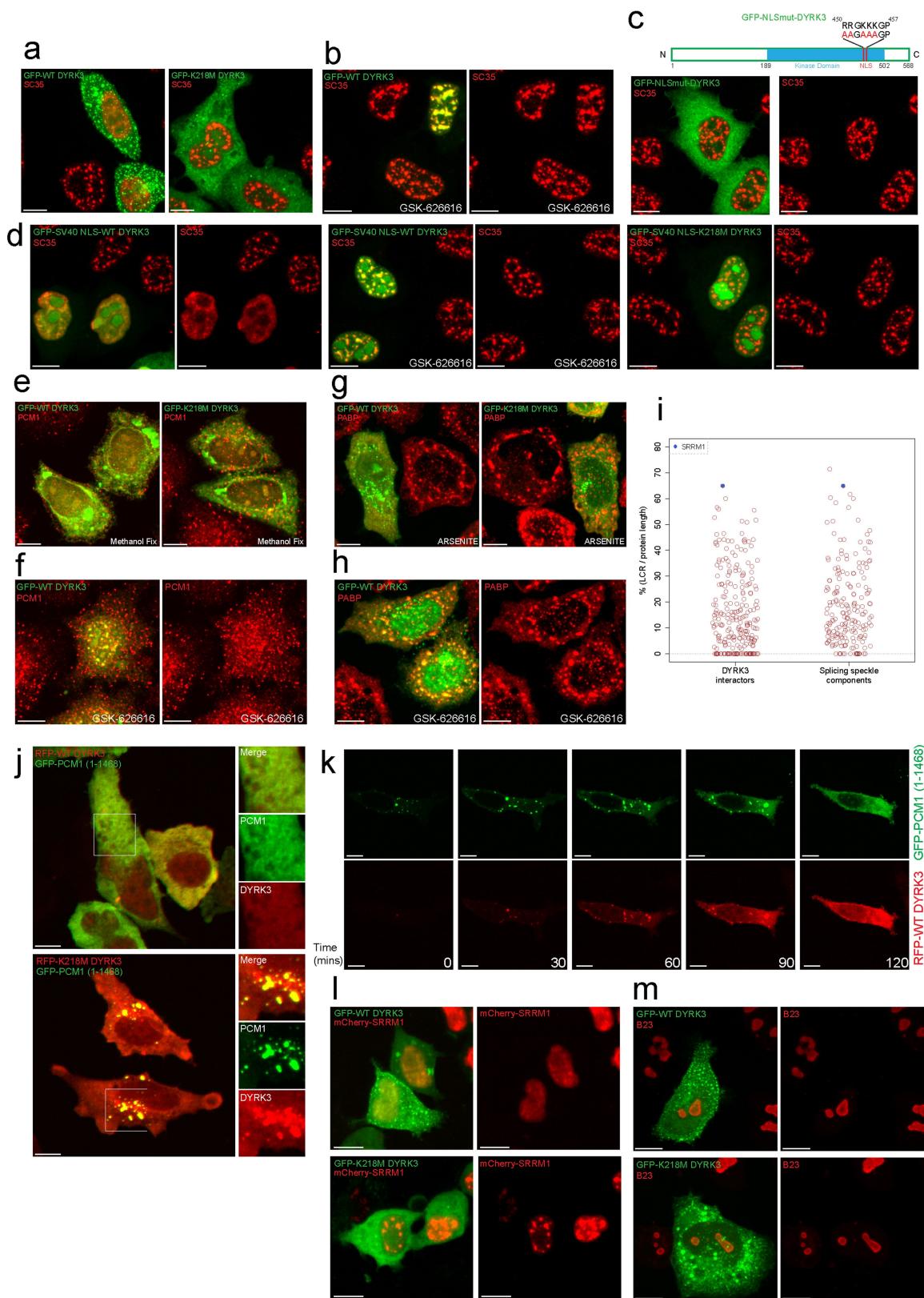
Extended Data Fig. 5 | Inhibition of DYRK3 does not affect all membraneless organelles in mitosis. **a**, Colocalization of splicing-speckle marker and poly(A) mRNA in hybrid compartments in mitotic cells upon GSK-626616 treatment (1 μ M, 3 h). **b**, Colocalization (arrowheads) of splicing-speckle marker and EGFP-DYRK3(WT) in hybrid compartments in mitotic cells upon GSK-626616 treatment (1 μ M, 2 h). **c**, Dissolution of P-bodies (DDX6) in mitotic cells is unaffected upon GSK-626616

treatment (1 μ M, 6 h). **d**, Dissolution of nucleoli (fibrillarin) in mitotic cells is unaffected upon GSK-626616 treatment (1 μ M, 6 h). **e**, Dissolution of Cajal bodies (coilin) in mitotic cells is unaffected upon GSK-626616 treatment (1 μ M, 6 hrs). **f**, Aberrant granules formed upon GSK-626616 treatment are not ubiquitinylated aggregates. Images are representative of at least three independent experiments. Scale bars, 10 μ m.



Extended Data Fig. 6 | Protein abundance does not change during the formation of mitotic granules upon GSK-626616 inhibitor treatment.
a, Time-lapse images (single Z-stack) show fusion of mitotic SRRM2-mCherry granules formed in the presence of the GSK-626616 inhibitor (1 μ M) in cells arrested in mitosis (thymidine–nocodazole block).
b, Mean cell intensity of SRRM2-mCherry is plotted during mitotic

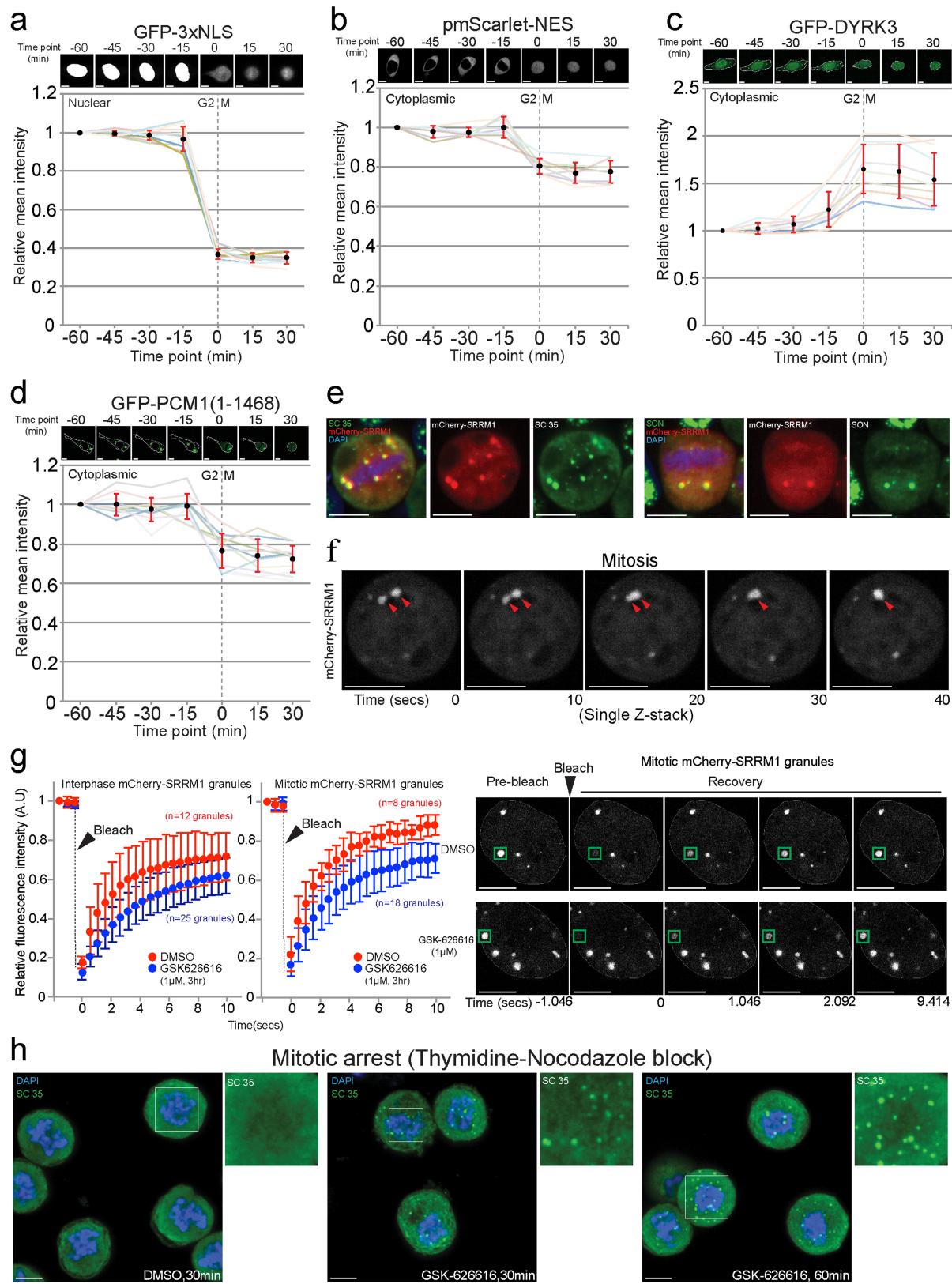
granule formation on addition of GSK-626616 inhibitor (1 μ M). Data are mean \pm s.d. **c**, Top, mean intensity of SRRM2-mCherry in the dissolved phase is plotted during mitotic granule formation on addition of GSK-626616 inhibitor (1 μ M). Bottom, time-lapse images of the cells plotted in the top panel. Images are representative of at least three independent experiments. Scale bars, 10 μ m.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Dissolution of membraneless organelles is dependent on DYRK3 localization and its kinase activity. **a**, Images show cells overexpressing wild-type EGFP–DYRK3(WT) and kinase-dead EGFP–DYRK3(K218M). These images are the same as Fig. 3a showing both channels. **b**, Dissolution of splicing speckles upon overexpression of EGFP–DYRK3(WT) in interphase cells is reversed by GSK-626616 treatment (1 μ M, 2 h). **c**, Top, schematic of the EGFP-tagged nuclear localization signal (NLS) mutant of DYRK3 (EGFP–NLSmut–DYRK3). Arginine and lysine residues in the NLS are mutated to alanine. Bottom, overexpressed EGFP–NLSmut–DYRK3 localizes to the cytoplasm and does not dissolve splicing speckles. **d**, Left, dissolution of splicing speckles upon overexpression of EGFP–NLS(SV40)–DYRK3(WT) in interphase cells. Middle, dissolution of splicing speckles upon overexpression of EGFP–NLS(SV40)–DYRK3(WT) in interphase cells is reversed upon GSK-626616 treatment (1 μ M, 2 h). Right, condensed splicing speckles upon overexpression of EGFP–NLS(SV40)–DYRK3(K218M) in interphase cells. **e**, Images show cells overexpressing EGFP–DYRK3(WT) and EGFP–DYRK3(K218M). These images are the same as Fig. 3c showing both channels. **f**, Dissolution of pericentriolar satellites upon overexpression

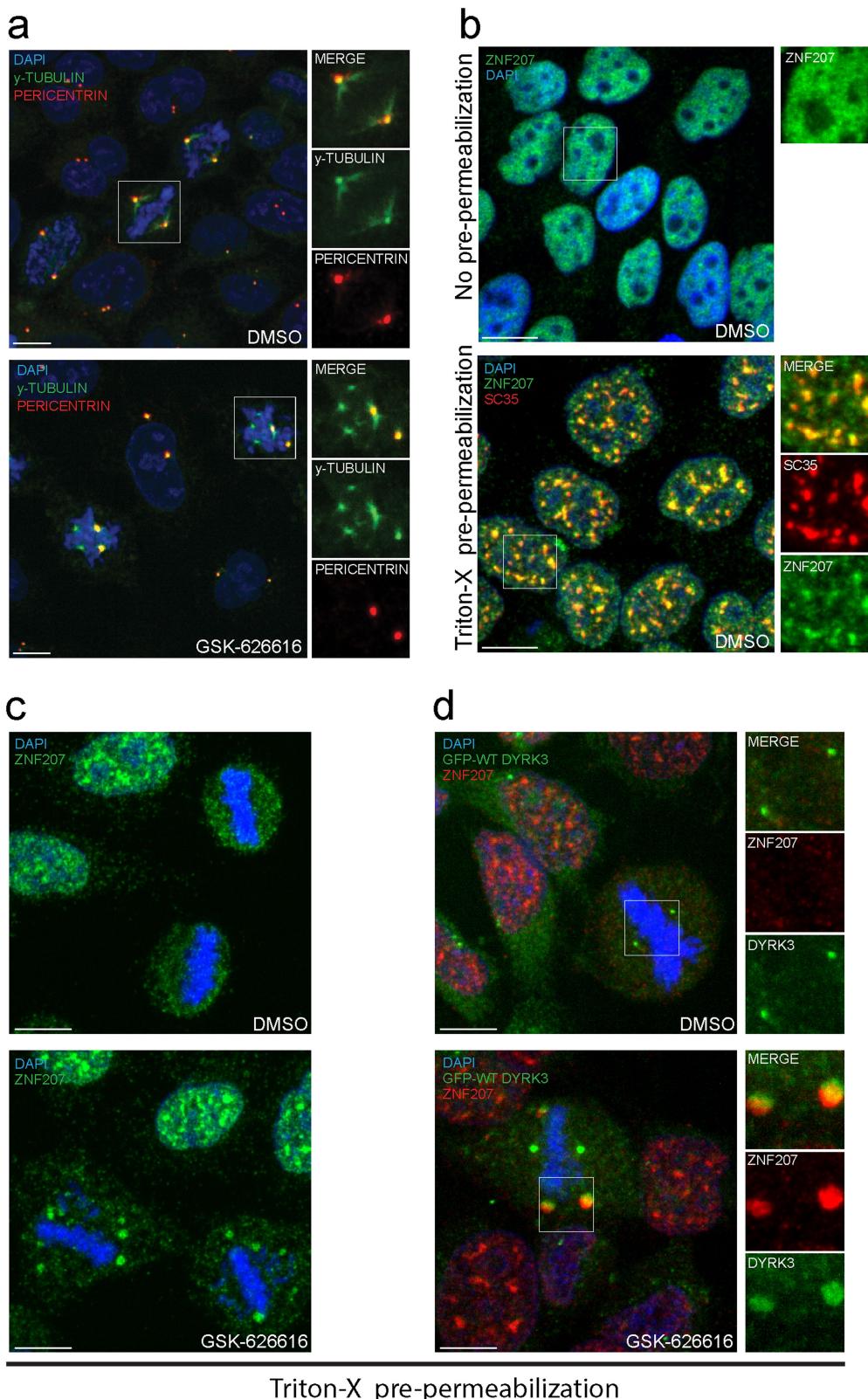
of EGFP–DYRK3(WT) in interphase cells is reversed upon GSK-626616 treatment (1 μ M, 2 h). **g**, Images show cells overexpressing EGFP–DYRK3(WT) and EGFP–DYRK3(K218M). These images are the same as Fig. 3e showing both channels. **h**, Dissolution of stress granules upon overexpression of EGFP–DYRK3(WT) in interphase cells is reversed upon GSK-626616 treatment (1 μ M, 2 h). **i**, The percentage of low complexity regions (LCR) occupying each full-length protein was computed for all DYRK3 interactors (Fig. 1a) and all the known splicing-speckle components (GO:0016607). SRRM1 shown in blue is among the proteins with the highest proportion of low complexity regions. **j**, Dissolution of EGFP–PCM1(1–1468) cytosolic granules upon overexpression of RFP–DYRK3(WT), and not RFP–DYRK3(K218M). **k**, Time-lapse images show RFP–DYRK3(WT) driven dissolution of cytosolic granules formed upon EGFP–PCM1(1–1468) overexpression. **l**, Dissolution of mCherry–SRRM1 nuclear granules upon overexpression of EGFP–DYRK3(WT), and not EGFP–DYRK3(K218M). **m**, Nucleoli are unaffected upon overexpression of EGFP–DYRK3(WT) or EGFP–DYRK3(K218M) in interphase cells. Images are representative of at least three independent experiments. Scale bars, 10 μ m.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Dilution of DYRK3 substrates during G2-to-M transition. **a**, Top, time-lapse images of a cell (single Z-stack) expressing EGFP-3 \times NLS during G2-to-M transition. Bottom, changes in mean nuclear intensity of EGFP-3 \times NLS during G2-to-M transition. Data are from eleven individual cells. **b**, Top, time-lapse images of a cell (single Z-stack) expressing pmScarlet-NES (nuclear export signal) during the G2-to-M transition. Bottom, changes in mean cytoplasmic intensity of pmScarlet-NES during the G2-to-M transition. Data are from eight individual cells. **c**, Top, time-lapse images of a cell (single Z-stack) expressing EGFP-DYRK3(WT) during G2-to-M transition. Bottom, changes in mean cytoplasmic intensity of EGFP-DYRK3(WT) during G2-to-M transition. Data are from nine individual cells. **d**, Top, Time-lapse images of a cell (single Z-stack) expressing EGFP-PCM1(1-1468) during G2-to-M transition. Bottom, changes in mean cytoplasmic intensity

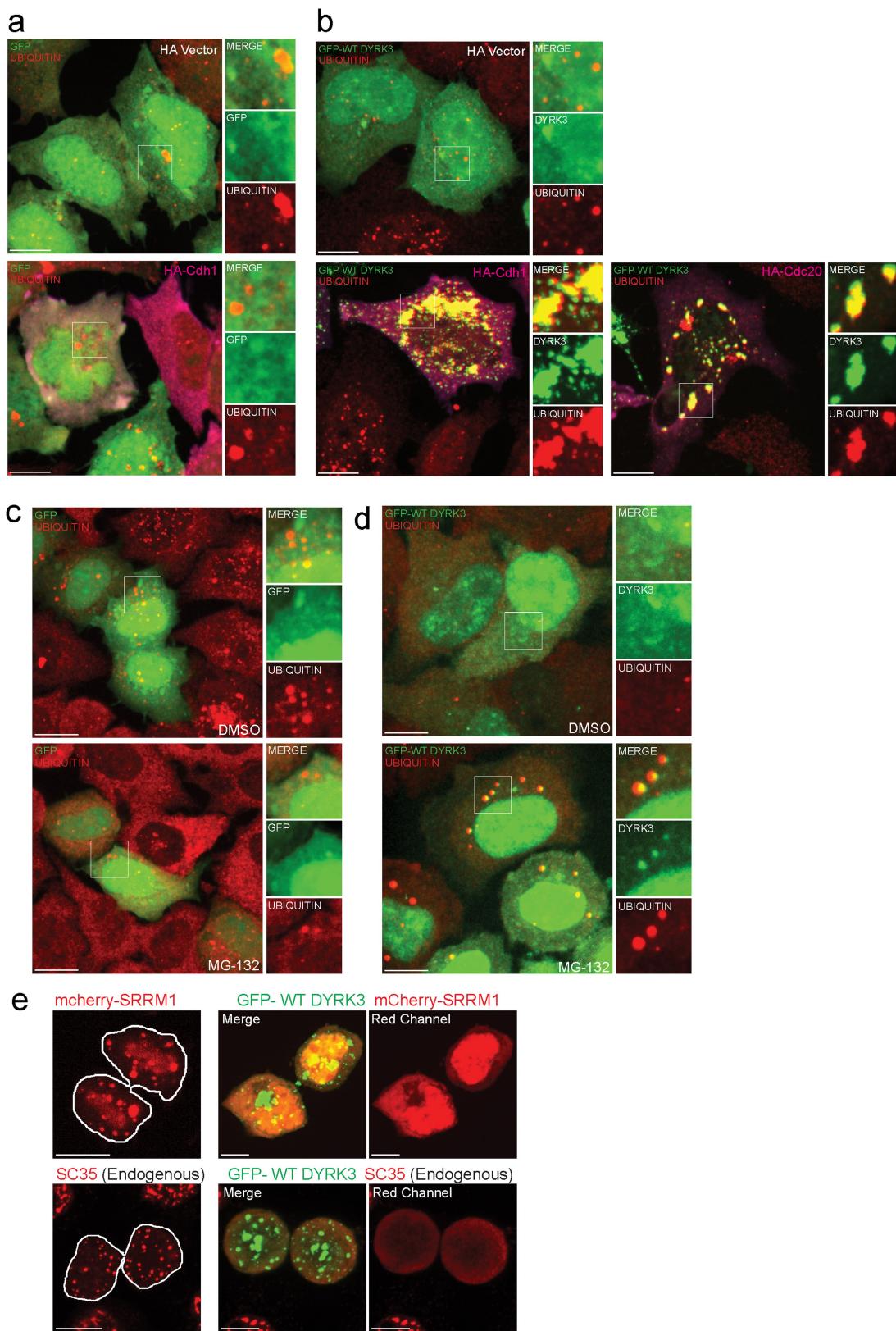
of EGFP-PCM1(1-1468) during G2-to-M transition. Data are from 11 individual cells. **a-d**, The lines (background) show mean nuclear intensity for individual cells. Data are mean \pm s.d. Time point (0 min) refers to nuclear envelope breakdown. **e**, Overexpressed mCherry-SRRM1 forms mitotic granules which recruit endogenous splicing proteins. **f**, Time-lapse images (single Z-stack) show fusion of mitotic mCherry-SRRM1 granules in mitotic cells. **g**, Left, FRAP analysis of interphase and mitotic mCherry-SRRM1 granules in the presence and absence of the GSK-626616 inhibitor (1 μ M). Data are mean \pm s.d. Right, FRAP recovery of mitotic mCherry-SRRM1 granule. **h**, Cells arrested in mitosis show formation of splicing granules upon GSK-626616 treatment (1 μ M, indicated times). Data and images are representative of at least three independent experiments. Scale bars, 10 μ m.



Triton-X pre-permeabilization

Extended Data Fig. 9 | Spindle apparatus defects upon DYRK3 inhibition. **a**, Multiple γ -tubulin foci in mitotic cells upon GSK-626616 treatment (1 μ M, 6 h). **b**, Top, ZNF207 localization in interphase cells (no pre-permeabilization). Bottom, ZNF207 and SC35 colocalize in Triton-X pre-permeabilized interphase cells. **c**, ZNF207 granules in mitotic cells upon GSK-626616 treatment (1 μ M, 3 h). **d**, EGFP-DYRK3(WT)

colocalizes with ZNF207 in mitotic cells upon GSK-626616 treatment (1 μ M, 3 h). EGFP-DYRK3(WT) expression was induced in HeLa-FlpIn-Trex cells by adding doxycycline (500 ng ml⁻¹, 6 h). Images are representative of at least three independent experiments. Scale bars, 10 μ m.



Extended Data Fig. 10 | DYRK3 forms ubiquitin-positive aggregates on overexpression of CDC20 and CDH1 or upon proteasomal inhibition. **a**, Ubiquitin does not colocalize with overexpressed EGFP inside cells. **b**, Ubiquitin localizes to EGFP-DYRK3(WT) aggregates, formed on HA-CDH1 and HA-CDC20 overexpression. **c**, Ubiquitin does not colocalize with overexpressed EGFP upon MG-132 treatment (5 μ M, 4 h). **d**, Ubiquitin localizes to EGFP-DYRK3(WT) granules upon

MG-132 treatment (5 μ M, 4 h). EGFP-DYRK3(WT) expression was induced in HeLa-FlpIn-Trex cells by adding doxycycline (500 ng ml $^{-1}$, 6 h). **e**, Top, EGFP-DYRK3(WT) overexpression prevents re-assembly of cytosolic mCherry-SRRM1 granules during late mitosis. Bottom, EGFP-DYRK3(WT) overexpression prevents re-assembly of cytosolic splicing granules during late mitosis. Images are representative of at least three independent experiments. Scale bars, 10 μ m.

Structure of the origin recognition complex bound to DNA replication origin

Ningning Li^{1,7}, Wai Hei Lam^{2,7}, Yuanliang Zhai^{2,3,6,7*}, Jiaxuan Cheng^{4,7}, Erchao Cheng⁴, Yongqian Zhao^{2,3}, Ning Gao^{1*} & Bik-Kwoon Tye^{2,5,*}

The six-subunit origin recognition complex (ORC) binds to DNA to mark the site for the initiation of replication in eukaryotes. Here we report a 3 Å cryo-electron microscopy structure of the *Saccharomyces cerevisiae* ORC bound to a 72-base-pair origin DNA sequence that contains the ARS consensus sequence (ACS) and the B1 element. The ORC encircles DNA through extensive interactions with both phosphate backbone and bases, and bends DNA at the ACS and B1 sites. Specific recognition of thymine residues in the ACS is carried out by a conserved basic amino acid motif of Orc1 in the minor groove, and by a species-specific helical insertion motif of Orc4 in the major groove. Moreover, similar insertions into major and minor grooves are also embedded in the B1 site by basic patch motifs from Orc2 and Orc5, respectively, to contact bases and to bend DNA. This work pinpoints a conserved role of ORC in modulating DNA structure to facilitate origin selection and helicase loading in eukaryotes.

Initiation of DNA replication begins with the binding of an initiator at a replicator followed by the recruitment of a replicative helicase that unwinds DNA^{1,2}. Bacterial and archaeal initiators contain an AAA+ domain and a helix-turn-helix (HTH) motif that are implicated in specific DNA interactions^{3–7}. In eukaryotes, the initiator is a highly conserved six-subunit origin recognition complex (ORC)^{8–10}. The five subunits Orc1–Orc5 each bear an AAA+ domain and a winged-helix domain (WHD)^{1,9}, whereas Orc6 bears little resemblance to the other ORC members and its role in the assembly of ORC around origin DNA seems to differ among organisms^{9,10}. Despite the high conservation of ORC in protein sequence among eukaryotes, its selectivity for replicators varies from species to species, in which specific DNA sequence has a determinant role in certain yeasts^{11,12} and chromatin structure has a predominant role in humans^{9,13}. Both of these modes are used in all eukaryotes but the prevalence of these selectivity modes is influenced by the interplay of several factors^{14–23}.

To understand how the ORC selects and binds DNA, much effort has been devoted to the study of the ORC architecture either alone or with Cdc6, which ORC recruits to origin DNA in G1 phase as a prelude to the assembly of the CMG (Cdc45–MCM–GINS) helicase^{24–29}. More recently, several high-resolution structures that include the ORC, such as the OCCM (ORC–Cdc6–Cdt1–Mcm2–Mcm7)³⁰, or that contain trimmed portions of ORC in fly³¹ and human³² have been determined. Together, these structures provide a clear picture of how ORC subunits interact with each other. Unfortunately, because DNA was either omitted or the resolution does not provide sufficient detail, we can only speculate on how the ORC may interact with DNA to perform its functions.

In this study, we have determined a series of structures of *S. cerevisiae* ORC, either alone or with ARS305 DNA (36 or 72 base pairs (bp) in length), using full-length proteins by single-particle cryo-electron microscopy (cryo-EM) (Extended Data Figs. 1, 2, Extended Data Table 1). The first ORC–DNA complex (36 bp, containing both ACS and B1 elements) was obtained at 3.6 Å, which allows a nearly complete identification of base sequence from end to end (Fig. 1a). However, the

interaction between the ORC and the B1 element is relatively unstable. Therefore, we extended the original DNA to 72 bp from the B1 end and determined a second structure of the ORC–DNA complex at 3.0 Å (Fig. 1b–d). These two structures are basically the same except that atomic interactions at the B1 proximal site in the second one are clearly resolved.

Subunit organization around the ACS DNA

In the ORC–DNA (72 bp) structure, over 50 bp of the 72-bp duplex, from one end to beyond the B1 element, can be visualized in the density map (Fig. 1b and Supplementary Video 1). The DNA beyond that point appears highly flexible. There is no melting of the duplex DNA, but a continuous curving of the backbone is clearly seen with two major bending points (Fig. 1d). One resides at position A8, right in the middle of ACS, and the other is in the middle of the B1 element at position T28.

Similar to previous structures^{30–32}, Orc1–Orc5 subunits oligomerise to encircle the ACS DNA through both canonical interactions between inter AAA+ domains (Fig. 1a–c and Extended Data Fig. 3a) and interdigitated domain-swapping interactions between the WHDs and the AAA+ domains of adjacent ORC subunits (Extended Data Fig. 3d). In addition to these conventional interactions, many flexible extensions and linkers of ORC subunits also contribute largely to the inter-subunit stabilization (Supplementary Video 1, Extended Data Fig. 3e–g). ORC recognizes and binds to origin DNA in an ATP-dependent manner⁸. Orc1–Orc5 subunits each contain AAA+ or AAA+-like domains, but only Orc1, Orc4 and Orc5 contain functional ATPase-related motifs^{9,33}. As expected, we identified three ATPγS molecules and their associated Mg²⁺ at three corresponding interfaces (Extended Data Fig. 4). However, each of the bound ATPγS molecules is uniquely coordinated, probably reflecting distinct roles for these centres during replication initiation (Supplementary Discussion).

Two regions of Orc6, the transcription factor II B (TFIIB) domain B (residues 271–386) and the carboxy-terminal domain (CTD, residues 387–430), were solved in the structure (Extended Data Fig. 1i). Orc6 interacts with three subunits, Orc3, Orc2 and Orc5, and their interfaces

¹State Key Laboratory of Membrane Biology, Peking-Tsinghua Center for Life Sciences, School of Life Sciences, Peking University, Beijing, China. ²Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong, China. ³Institute of Advanced Study, The Hong Kong University of Science and Technology, Hong Kong, China. ⁴Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China. ⁵Department of Molecular Biology & Genetics, Cornell University, Ithaca, NY, USA. ⁶Present address: School of Biological Sciences, The University of Hong Kong, Hong Kong, China. ⁷These authors contributed equally: Ningning Li, Wai Hei Lam, Yuanliang Zhai, Jiaxuan Cheng. *e-mail: zhai@hku.hk; gaon@pku.edu.cn; biktye@ust.hk

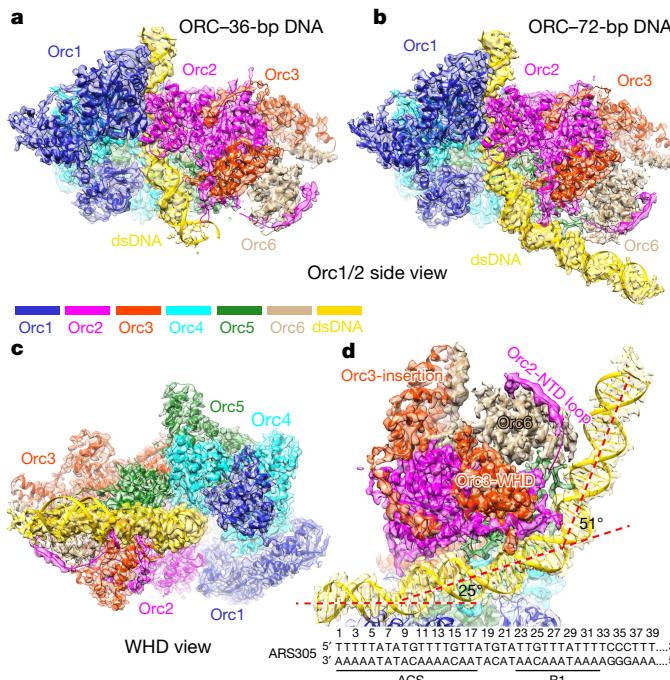


Fig. 1 | Overall structure of ORC bound to origin DNA. **a**, Side view of the cryo-EM density map of the ORC–DNA complex (36 bp) with the atomic model superimposed. dsDNA, double-stranded DNA. **b**, **c**, Side (Orc1/Orc2) (**b**) and bottom (WHD) (**c**) views of the ORC–DNA complex (72 bp). **d**, Cut-away view of the ORC–72-bp DNA complex, highlighting the interaction of Orc6 with Orc2 and Orc3. ARS305 sequence is shown, with successive bending of the DNA illustrated by the dashed red lines. Other Orc subunits are omitted for clarity. NTD, N-terminal domain.

are dominated by hydrophobic residues (Extended Data Fig. 5a–d). Orc6 is situated distal to the ORC central channel and is not involved in the ACS DNA recognition but interacts with the B1 element through its TFIIB domain (Fig. 1d, Extended Data Fig. 5e).

The AAA+ modules of the Orc1–Orc5 subunits form a tilted and breached ring around DNA with a gap between Orc1 and Orc2 (Extended Data Fig. 3a). Notably, the Orc2 WHD is sandwiched between the AAA+ domains of Orc1 and Orc2, deviating from the WHD tier (Extended Data Fig. 3a, b). This is in contrast to the OCCM, in which the six WHDs occupy nearly symmetric positions³⁰. 3D classification of the ORC–DNA particles showed that the Orc2 WHD is indeed flexible and takes up distinct positions in different subpopulations (Extended Data Fig. 6a–c). In fact, before DNA binding, as seen in our structure of ORC alone, both the entire Orc1-AAA+ and Orc2-WHD appear highly flexible and a very large opening can be observed between Orc1 and Orc2 (Extended Data Fig. 6h). This observation is consistent with the low-resolution electron microscopy structure of human ORC1–ORC5, in which a gap was also observed between ORC1 and ORC2³². Also consistently, the crystal structure of *Drosophila* ORC³¹ in an auto-inhibitory state showed that Orc1-AAA+ and Orc2-WHD occupy completely different positions (Extended Data Fig. 7). The function of ORC in helicase loading requires Cdc6 to be inserted transiently into the gap between Orc1 and Orc2^{30,34}. Therefore, these structures indicate that the interface between Orc1 and Orc2 is intrinsically dynamic, serving for both DNA entry and Cdc6 dock.

Non-specific gripping of origin DNA by ORC

The archaeal initiator Orc1 recognizes origin DNA using three major modules, an initiation-specific motif (ISM) from the AAA+ domain, a β-hairpin ‘wing’ and one HTH motif from the WHD^{5,6}. These DNA recognition modules are largely preserved in each of the Orc1–Orc5 subunits in eukaryotes; however, their roles in contacting DNA have diverged. In the ORC–DNA structure, the five ISMs are arranged in a helical spiral around DNA, but Orc4-ISM has no direct contact with

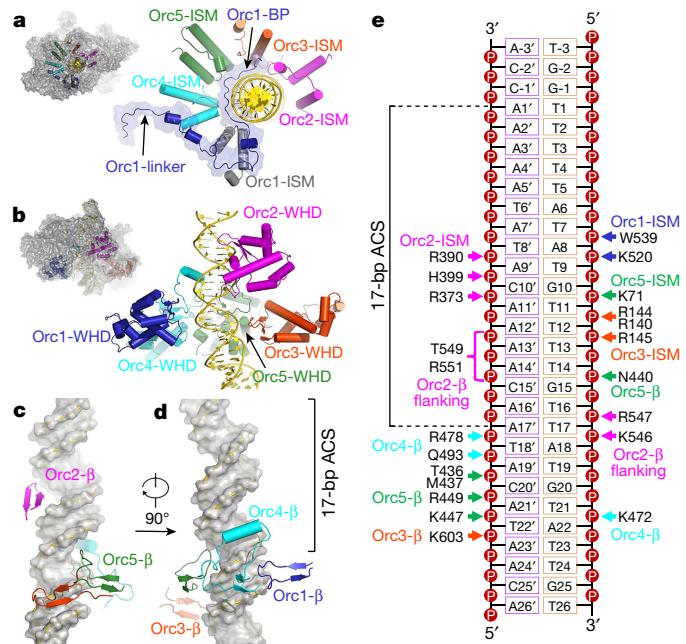


Fig. 2 | Extensive interactions between ORC and DNA around the ACS region. **a**, Distribution of the ISMs of ORC subunits around origin DNA. The long N-terminal loop upstream of Orc1-AAA+ is highlighted in blue. Orc1-BP is inserted between the ISMs of Orc1, Orc4, Orc5, Orc3 and DNA. Orientations of the ISMs in the ORC–DNA structure are shown in top-left thumbnail. **b**, Same as in **a**, but for the five WHDs of the ORC subunits. **c**, **d**, Distribution of the β-hairpin motifs from the WHDs of ORC subunits around the DNA (surface representation). **e**, Summary of non-specific ORC–DNA backbone interactions around the ACS region. The residues of ORC subunits, mostly basic, that contribute to DNA backbone phosphate interactions are labelled.

origin DNA (Fig. 2a, e). In sharp contrast to the archaeal ORC, all of the HTH motifs of Orc1–Orc5 are positioned away from DNA (Extended Data Fig. 3c). Three β-hairpin motifs of the WHDs (Fig. 2b) from Orc2, Orc4 and Orc5 contact the major groove in a helical manner, spanning nearly a full turn of the ACS DNA (Fig. 2c, d). Particularly, β-hairpins of Orc2 and Orc5 are both inserted into the major groove, with either the loop or flanking residues interacting with DNA backbone (Fig. 2c–e). The β-hairpin of Orc3, although not inserted into the major groove, is also in close contact with DNA phosphate backbone (Fig. 2c–e). Notably, some of the DNA contacts through these conserved modules of ORC are dynamic. As seen in the OCCM structure³⁰, ISMs of Orc1 and Orc5 and the β-hairpin of Orc2 are no longer in contact with DNA. Together, these features suggest that these conserved elements of ORC act combinatorially, probably through a universal mechanism, to provide a versatile grip for holding origin DNA with fairly high affinity but low sequence-specificity in different functional states of ORC.

Specific recognition of ACS by Orc1, Orc2 and Orc4

For specific origin DNA binding, eukaryotic ORC has acquired new elements to augment the cross-kingdom conserved DNA binding modules. Through structural analysis, we identified three critical motifs from Orc1, Orc2 and Orc4 that are important for sequence-specific recognition of ACS DNA. Notably, a basic patch of Orc1 (residues 358–371, Orc1-BP), which is located in a highly disordered region between the BAH and the AAA+ domain, is fully inserted into the minor groove of the ACS DNA for a half turn atop of all the other DNA-binding modules (Figs. 2a, 3a, b). This insertion is facilitated by Orc2-ISM, which places a bulky Trp396 deep into the minor groove to form a hydrogen bond between the N1 of the indole group and the O2 of the thymine ring from the invariant T11 of the ACS (Fig. 3c). This tryptophan insertion seems to serve as a steric roadblock, forcing the upstream sequence of Orc1-BP to make a 90° turn away from DNA (Fig. 3a).

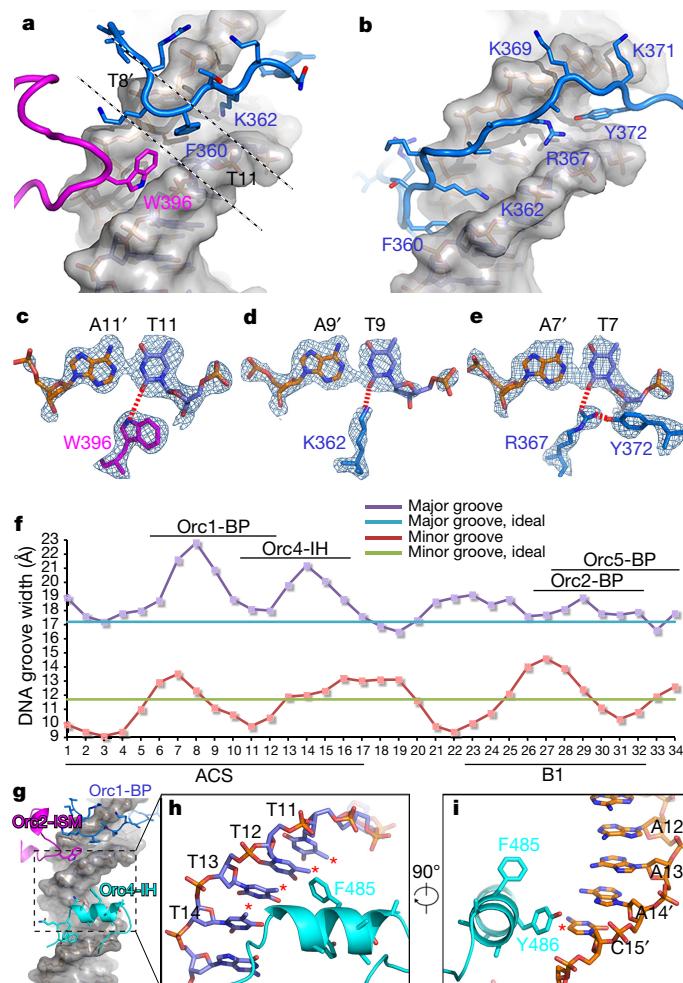


Fig. 3 | Base-specific recognition of ACS DNA by ORC. **a**, Zoomed-in view showing insertion of Orc1-BP into the minor groove of the ACS. Orc2-ISM and Orc1-BP are coloured magenta and blue, respectively. DNA is highlighted in grey surface representation. Selected residues from Orc1 and Orc2 and relevant nucleotides are labelled. **b**, Same as in **a**, but in a 90°-rotated view. **c**, Base recognition of T11 from the T-rich strand by W396 of Orc2. The DNA base pair is shown in stick model with cryo-EM density superimposed. The hydrogen bond between W396 and O2 of T11 is indicated by dashed lines (approximately 3.0 Å). Prime symbols denote bases on the opposite strand. **d**, Same as **c**, but for the hydrogen bond (approximately 2.5 Å) between K362 and the O2 of T9 (T-rich strand). **e**, Same as in **c**, but for the hydrogen bonds (approximately 2.5 Å) between R367 and the O2 of T7 (T-rich strand) as well as the OH of Y372. **f**, Plots

of DNA groove widths of the ORC-bound DNA. Widths were measured as the distances between opposite phosphates of major and minor grooves. The major and minor groove widths of ideal B-form DNA are shown as two constant values, 17 and 12 Å, respectively. Contact regions for groove-inserting motifs are marked. Measurements were done with the program 3DNA⁴⁹. **g**, The major groove insertion of the ACS by Orc4-IH. Minor groove insertion by Orc1-BP is also shown. **h**, Magnified view of the boxed region in **g** showing the hydrophobic interaction between F485 and the T-stretch (T11-T14) of the T-rich strand. Red asterisks denote the methyl group of thymine. **i**, Same as in **h**, but in a 90°-rotated view for the hydrophobic interaction between Y486 of Orc4-IH and C15' of the A-rich strand. The red asterisk denotes the Hoogsteen edge of C15'.

Upon binding of Orc1-BP, the major groove is widened by more than 4 Å while the minor groove is widened by 2.5 Å at the C-terminal half but compressed by more than 1 Å at the N-terminal end of the basic patch, relative to ideal B-form DNA (Fig. 3f). By embedding in the minor groove, the side chains of Lys362 and Arg367 establish specific interactions with the bases of T9 and T7, respectively, on the T-rich strand via hydrogen bonds (Fig. 3d, e). These two specific thymine recognitions are facilitated by flanking residues near Lys362 and Arg367. The guanidinium group of Arg367 coordinates with the hydroxyl group of Tyr372 (also inserted in the minor groove) and O2 of T7 in a perfect triangular configuration (Fig. 3e). Phe360 (also inserted into the minor groove) is sandwiched between two opposing deoxyribose moieties of opposite strands by hydrophobic interactions (Fig. 3a), probably contributing to the positioning of the side chain of Lys362 towards the O2 of T9 (Fig. 3d).

Importantly, this basic patch was proposed as the eukaryotic origin sensor, as substitution mutations of Lys362 and Arg367 showed that they are crucial in both *in vitro* ACS binding and cell viability³⁵. Basic

patches are common motifs among DNA-binding proteins such as TATA box binding proteins and high mobility group (HMG) proteins that show a preference for A-T bases³⁶. Comparative sequence analyses show that similar basic patches are found in Orc1 across species from yeast to humans³⁵ (Extended Data Fig. 8a), suggesting that this motif may also help to anchor ORC in T-rich sequences in all eukaryotes. In support of this hypothesis, metazoan ORCs from frog, fly and human also prefer AT-rich DNA substrates to some extent^{37–41}.

Another important motif contributing to the sequence-specific interaction is a special insertion α -helix (IH) in the β -hairpin of Orc4-WHD. As also seen in OCCM³⁰, Orc4-IH is deeply inserted into the major groove of the ACS (Fig. 3g). In particular, this region contains four consecutive thymines (T11–T14) in ARS305, and their methyl groups create a local hydrophobic environment (Fig. 3h). Orc4-IH, on the other hand, contains mostly hydrophobic residues with four aromatic side chains, in which Phe485 is especially close to the methyl groups of T12 and T13 (Fig. 3h). This observation suggests that Orc4-IH is most likely involved in sequence-specific

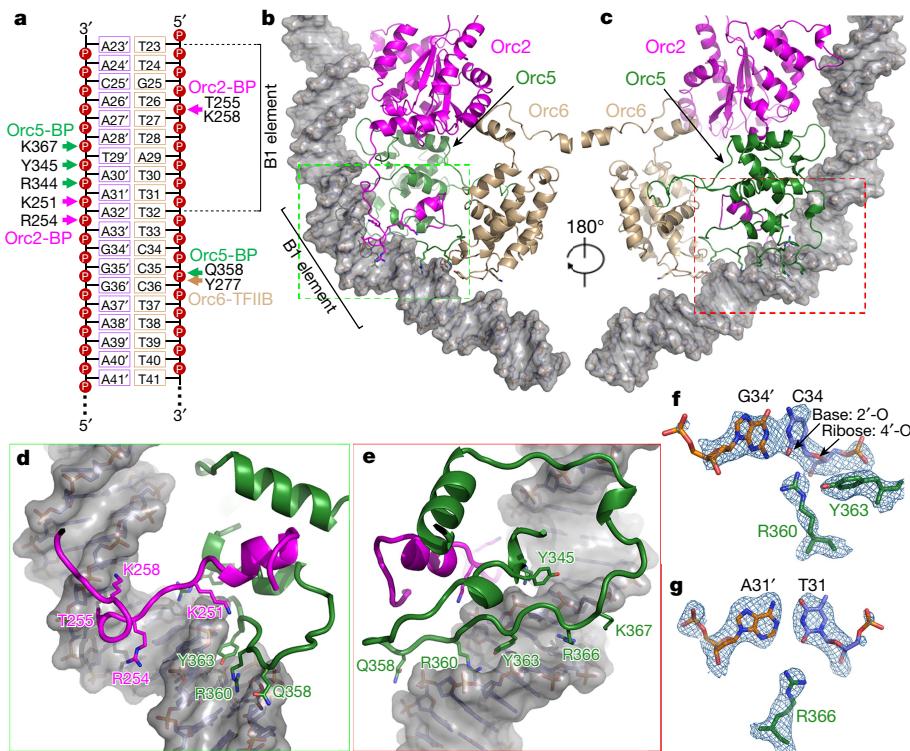


Fig. 4 | Interactions between ORC and DNA around the B1 element.

a, Summary of non-specific ORC–DNA backbone interactions around B1 region. **b, c**, Orc2, Orc5, and Orc6 interact with DNA around the B1 site. **d**, Magnified view of the boxed region in **b** showing insertion of the helical turn motif of Orc2 in the major groove. **e**, Magnified view of the boxed

region in **c** showing insertion of Orc5-BP in the minor groove. **f**, Arg360 of Orc5-BP is orientated towards the O2 of C34 and O4 of the sugar ring (both less than 3 Å distance). **g**, Arg366 of Orc5-BP points to the O2 of T31 of the T-rich strand (approximately 3.5 Å).

interaction with the thymine stretch in the ACS sequence. Upon Orc4-IH binding, the major groove is widened by 4 Å compared to ideal B-form DNA (Fig. 3f). In addition, the hydrophobic ring of Tyr486 stacks upon the hydrophobic Hoogsteen edge of the conserved C15' from the A-rich strand (Fig. 3i), also contributing to sequence specificity. Notably, this special insertion α -helix is completely absent in metazoans (Extended Data Fig. 8b), which may partially explain the divergent specificity of ORC for origin DNA among different species³⁰.

Throughout these analyses, we have discovered several base-specific interactions from both the conserved and species-specific elements of the ORC subunits. Notably, most of them involve thymine nucleotides through specific recognition of the methyl group from the major groove (Hoogsteen edge) or the free O2 from the minor groove (sugar edge) of the thymine ring (Fig. 3). These thymine-specific interactions perfectly explain the ‘uniqueness’ in structure and sequence of the signature asymmetric T-rich sequence of the *S. cerevisiae* ACS.

B1 interaction with Orc2, Orc5 and Orc6

Biochemical and genetic evidence suggest that ORC also binds to the B1 element of origins^{24,42,43}. Notably, in the ORC–DNA structure, two motifs from Orc2 and Orc5 were found to be inserted in the major and minor grooves of the B1 DNA, respectively (Fig. 4). A flexible loop (residues 251–258, also a basic patch) upstream Orc2-AAA+ forms a helical turn in the major groove (Fig. 4b, d) contacting DNA backbone through basic residues such as Lys251, Arg254 and Lys258 (Fig. 4a, d). In analogy to the ACS recognition by Orc1-BP, a basic patch (residues 358–367) from Orc5-WHD is inserted into the adjacent minor groove (Fig. 4c, e), which is also conserved to a certain extent across species (Extended Data Fig. 8c). However, by contrast, these embedded side chains of Orc5-BP display relatively weak interactions with the bases. Arg360 of Orc5-BP, coordinated by Tyr363, points to the O2 of C34 and the O4 of the sugar ring (both approximately 3 Å); Arg366 is orientated towards the O2 of T31, at a distance of around 3.5 Å (Fig. 4f, g). As

to Orc6, the N-terminal sequence of its TFIIB domain B (TFIIB-B), including Tyr277, interacts with the backbone of DNA (Fig. 4a, Extended Data Fig. 5e). Importantly, as seen in the structures from 3D classification, the extent of DNA curving positively correlates with the stability of Orc6 in the ORC–DNA complex (Extended Data Fig. 6f, g), suggesting the bending at the B1 site comes from cooperative effects from all three interacting ORC subunits.

The lack of a strong discrimination of bases at the B1 interaction site echoes the diversified B1 elements in yeast origins⁴⁴. The A/T-rich nature of B1 element suggests that Orc5-BP could potentially contribute to specific base recognition as well, because the DNA bending is a dynamic process and hydrogen bonds similar to Orc1-BP could be established readily with the O2 of thymine in the minor groove of the B1 element.

Implications in origin DNA selection and MCM loading

There are three salient lessons learned from this study. First is the importance of Orc1-BP in binding to the minor groove of the ACS to recognize three conserved bases, T7, T9 and T11 (Fig. 3). We note that there is some degeneracy in the T7 and T9 positions for either A or T in the ACS (Fig. 5a). An explanation for this degeneracy is that hydrogen donors such as Arg367 are in the same proximity to hydrogen receptors of O2 of thymine or N3 of adenine in the minor groove such that A–T or T–A base pairs are indistinguishable for hydrogen bonding^{45,46}. Another explanation is that induced-fit by the flexible Orc1-BP allows contact with the O2 of thymine on either strand at these positions. Further examination of Orc1 sequence indicates that multiple basic patches are present in the flexible linker region (Extended Data Fig. 8e). Notably, these additional basic patches in Orc1 are also found in other species (Extended Data Fig. 8e, f).

Second is the role of Orc4-IH in binding to consecutive thymines (11–14, mostly invariant) and a conserved C15' (A-rich strand) of the ACS in the major groove (Fig. 3g–i). Given the fact that this Orc4-IH is absent in metazoans (Extended Data Fig. 8b), Orc1-BP may be the

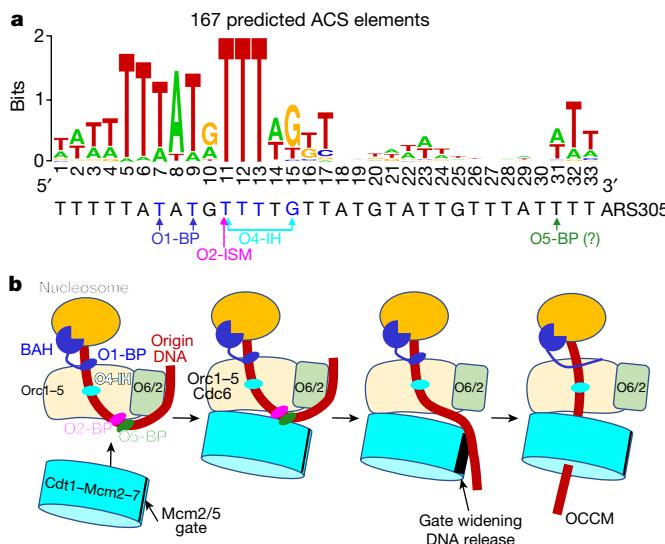


Fig. 5 | Model of origin recognition and MCM loading by ORC. **a**, Position weight matrix of the ORC binding consensus built with a predicted set including 167 ACS elements. All ACS used in these analyses were previously published¹⁶ (Supplementary Table 1). The sequence logo was generated using <http://weblogo.berkeley.edu>. Conserved bases of ARS305 recognized by ORC motifs (O1-BP, O2-ISM, O4-IH and probably O5-BP) are highlighted. **b**, Schematic model showing the positioning of ORC next to a nucleosome and the subsequent loading of MCM. Docking of ORC at the initiation site is facilitated by the binding of Orc1-BAH to the nucleosome and the searching of ACS by Orc1-BPs. Establishment of ACS and B1 interactions by ORC base-specific recognition motifs curves the DNA to facilitate docking of Cdt1-MCM. Regulated DNA liberation and widening of the Mcm2/Mcm5 gate ensure precise insertion of origin DNA into the MCM ring.

major or sole determinant for DNA recognition and hence the more relaxed sequence specificity for ORC in metazoans.

Third is the bending of the origin DNA at successive points at the ACS and B1 elements (Fig. 1d) induced by deformation of major and minor grooves through interactions with various ORC subunits (Fig. 3f). As a consequence of this bending, the TFIIB-B of Orc6 conceivably could come into contact with the origin DNA beyond the B1 element (Extended Data Fig. 5g). This particular feature, if true, will shed light on the special nature of ORC–origin DNA complex in organizing chromatin structure. The formation of a nucleosome-like structure between ORC and origin DNA has been postulated based on biochemical studies^{24,25} and a low-resolution electron microscopy study of *Drosophila* ORC bound to DNA²⁹. This nucleoprotein structure has been implicated to provide a nucleosome-free region for replication initiation by phasing adjacent nucleosomes^{15,16,21}.

In brief, our ORC–DNA structure provides important insights into origin recognition in all eukaryotes. Our working hypothesis is that ORC-binding sites in eukaryotes are determined by a combination of factors at least including the BAH domain and Orc1-BP. The BAH domain of yeast, fly, plant and human Orc1 are known to interact with nucleosomes juxtaposed to replication origins^{9,14,20,23}. We believe that the BAH domain and Orc1 basic patches may be involved in the initial searching and anchoring of ORC at DNA sites close to the designated initiation site (Fig. 5b). The precise binding to specific sites for replication initiation is likely to depend on a variety of means such as recruitment by other protein factors^{14,47}, modified histones, and specific DNA structures (G4 DNA)⁴⁸. In *S. cerevisiae*, the positioning of ORC at the ACS is achieved by specific interactions between the essential Orc1-BP and Orc4-IH with invariant bases of the ACS (Fig. 5b).

Although the specificity of ORC in DNA binding is highly diverged, DNA bending induced by ORC could be shared by all eukaryotes to serve as a mechanism for the loading of MCM complex by ORC onto origin DNA. Notably, superimposition of the ORC–DNA and

OCCM structures³⁰ indicates that the bent DNA is positioned at the gate between the Mcm2 and Mcm5 subunits of the MCM ring in the OCCM (Extended Data Fig. 9). This perfect alignment suggests that DNA bending by ORC has a unique function in MCM loading. We envision that by bending, DNA is tugged away from the interacting surfaces of ORC and the MCM ring to maximize their contacts while aligned with the Mcm2/Mcm5 gate of the MCM ring (Fig. 5b). The coordination of DNA straightening and gate opening in the formation of OCCM is a testable hypothesis.

In summary, information derived from the ORC–DNA structure not only provides a structural framework for understanding how ORC recognizes and binds yeast origin DNA, but also provides insight into the origin selection mechanism in metazoans. The induced bending of origin DNA paves the way for studying the potentially conserved roles of ORC in regulating both MCM loading and chromatin organization in all eukaryotes.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0293-x>.

Received: 26 January 2018; Accepted: 8 May 2018;

Published online 4 July 2018.

1. Bleichert, F., Botchan, M. R. & Berger, J. M. Mechanisms for initiating cellular DNA replication. *Science* **355**, eaah6317 (2017).
2. Costa, A., Hood, I. V. & Berger, J. M. Mechanisms for initiating cellular DNA replication. *Annu. Rev. Biochem.* **82**, 25–54 (2013).
3. Erzberger, J. P., Mott, M. L. & Berger, J. M. Structural basis for ATP-dependent DNA assembly and replication-origin remodeling. *Nat. Struct. Mol. Biol.* **13**, 676–683 (2006).
4. Duderstadt, K. E., Chuang, K. & Berger, J. M. DNA stretching by bacterial initiators promotes replication origin opening. *Nature* **478**, 209–213 (2011).
5. Gaudier, M., Schuwirth, B. S., Westcott, S. L. & Wigley, D. B. Structural basis of DNA replication origin recognition by an ORC protein. *Science* **317**, 1213–1216 (2007).
6. Dueber, E. L., Corn, J. E., Bell, S. D. & Berger, J. M. Replication origin recognition and deformation by a heterodimeric archaeal Orc1 complex. *Science* **317**, 1210–1213 (2007).
7. Miller, J. M. & Enemark, E. J. Fundamental Characteristics of AAA+ Protein Family Structure and Function. *Archaea* **2016**, 9294307 (2016).
8. Bell, S. P. The origin recognition complex: from simple origins to complex functions. *Genes Dev.* **16**, 659–672 (2002).
9. Hoggard, T. & Fox, C. A. in *The Initiation of DNA Replication in Eukaryotes* (ed. Kaplan, D. L.) 159–188 (Springer International Publishing, 2016).
10. Duncker, B. P., Chesnokov, I. N. & McConkey, B. J. The origin recognition complex protein family. *Genome Biol.* **10**, 214 (2009).
11. Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**, 128–134 (1992).
12. Nieduszynski, C. A., Knox, Y. & Donaldson, A. D. Genome-wide identification of replication origins in yeast by comparative genomics. *Genes Dev.* **20**, 1874–1879 (2006).
13. Miotto, B., Ji, Z. & Struhl, K. Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proc. Natl. Acad. Sci. USA* **113**, E4810–E4819 (2016).
14. Müller, P. et al. The conserved bromo-adjacent homology domain of yeast Orc1 functions in the selection of DNA replication origins within chromatin. *Genes Dev.* **24**, 1418–1433 (2010).
15. Lipford, J. R. & Bell, S. P. Nucleosomes positioned by ORC facilitate the initiation of DNA replication. *Mol. Cell* **7**, 21–30 (2001).
16. Eaton, M. L., Galani, K., Kang, S., Bell, S. P. & MacAlpine, D. M. Conserved nucleosome positioning defines replication origins. *Genes Dev.* **24**, 748–753 (2010).
17. Deshpande, A. M. & Newlon, C. S. The ARS consensus sequence is required for chromosomal origin function in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **12**, 4305–4313 (1992).
18. Rao, H., Marahrens, Y. & Stillman, B. Functional conservation of multiple elements in yeast chromosomal replicators. *Mol. Cell. Biol.* **14**, 7643–7651 (1994).
19. Royzman, I., Austin, R. J., Bosco, G., Bell, S. P. & Orr-Weaver, T. L. ORC localization in *Drosophila* follicle cells and the effects of mutations in dE2F and dDP. *Genes Dev.* **13**, 827–840 (1999).
20. Kuo, A. J. et al. The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* **484**, 115–119 (2012).
21. Berbenetz, N. M., Nislow, C. & Brown, G. W. Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genet.* **6**, e1001092 (2010).

22. Simpson, R. T. Nucleosome positioning can affect the function of a cis-acting DNA element *in vivo*. *Nature* **343**, 387–389 (1990).
23. Li, S. et al. Structural basis for the unique multivalent readout of unmodified H3 tail by *Arabidopsis* ORC1b BAH-PHD cassette. *Structure* **24**, 486–494 (2016).
24. Lee, D. G. & Bell, S. P. Architecture of the yeast origin recognition complex bound to origins of DNA replication. *Mol. Cell. Biol.* **17**, 7159–7168 (1997).
25. Speck, C., Chen, Z., Li, H. & Stillman, B. ATPase-dependent cooperative binding of ORC and Cdc6 to origin DNA. *Nat. Struct. Mol. Biol.* **12**, 965–971 (2005).
26. Chen, Z. et al. The architecture of the DNA replication origin recognition complex in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **105**, 10326–10331 (2008).
27. Sun, J. et al. Cdc6-induced conformational changes in ORC bound to origin DNA revealed by cryo-electron microscopy. *Structure* **20**, 534–544 (2012).
28. Bleichert, F. et al. A Meier–Gorlin syndrome mutation in a conserved C-terminal helix of Orc6 impedes origin recognition complex formation. *eLife* **2**, e00882 (2013).
29. Clarey, M. G., Botchan, M. & Nogales, E. Single particle EM studies of the *Drosophila melanogaster* origin recognition complex and evidence for DNA wrapping. *J. Struct. Biol.* **164**, 241–249 (2008).
30. Yuan, Z. et al. Structural basis of Mcm2–7 replicative helicase loading by ORC–Cdc6 and Cdt1. *Nat. Struct. Mol. Biol.* **24**, 316–324 (2017).
31. Bleichert, F., Botchan, M. R. & Berger, J. M. Crystal structure of the eukaryotic origin recognition complex. *Nature* **519**, 321–326 (2015).
32. Tocilj, A. et al. Structure of the active form of human origin recognition complex and its ATPase motor module. *eLife* **6**, e20818 (2017).
33. Klemm, R. D., Austin, R. J. & Bell, S. P. Coordinate binding of ATP and origin DNA regulates the ATPase activity of the origin recognition complex. *Cell* **88**, 493–502 (1997).
34. Zhai, Y. et al. Unique roles of the non-identical MCM subunits in DNA replication licensing. *Mol. Cell* **67**, 168–179 (2017).
35. Kawakami, H., Ohashi, E., Kanamoto, S., Tsurimoto, T. & Katayama, T. Specific binding of eukaryotic ORC to DNA replication origins depends on highly conserved basic residues. *Sci. Rep.* **5**, 14929 (2015).
36. Bewley, C. A., Gronenborn, A. M. & Clore, G. M. Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 105–131 (1998).
37. Austin, R. J., Orr-Weaver, T. L. & Bell, S. P. *Drosophila* ORC specifically binds to ACE3, an origin of DNA replication control element. *Genes Dev.* **13**, 2639–2649 (1999).
38. Chesnokov, I., Remus, D. & Botchan, M. Functional analysis of mutant and wild-type *Drosophila* origin recognition complex. *Proc. Natl Acad. Sci. USA* **98**, 11997–12002 (2001).
39. Kong, D., Coleman, T. R. & DePamphilis, M. L. *Xenopus* origin recognition complex (ORC) initiates DNA replication preferentially at sequences targeted by *Schizosaccharomyces pombe* ORC. *EMBO J.* **22**, 3441–3450 (2003).
40. Vashee, S. et al. Sequence-independent DNA binding and replication initiation by the human origin recognition complex. *Genes Dev.* **17**, 1894–1908 (2003).
41. Liu, J. et al. DNA sequence templates adjacent nucleosome and ORC sites at gene amplification origins in *Drosophila*. *Nucleic Acids Res.* **43**, 8746–8761 (2015).
42. Rao, H. & Stillman, B. The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. *Proc. Natl Acad. Sci. USA* **92**, 2224–2228 (1995).
43. Rowley, A., Cocker, J. H., Harwood, J. & Diffley, J. F. Initiation complex assembly at budding yeast replication origins begins with the recognition of a bipartite sequence by limiting amounts of the initiator, ORC. *EMBO J.* **14**, 2631–2641 (1995).
44. Lucas, I. A. & Raghuraman, M. K. The dynamics of chromosome replication in yeast. *Curr. Top. Dev. Biol.* **55**, 1–73 (2003).
45. Hélène, C. DNA recognition. Reading the minor groove. *Nature* **391**, 436–438 (1998).
46. Seeman, N. C., Rosenberg, J. M. & Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA* **73**, 804–808 (1976).
47. Thomae, A. W. et al. Interaction between HMGA1a and the origin recognition complex creates site-specific replication origins. *Proc. Natl Acad. Sci. USA* **105**, 1692–1697 (2008).
48. Hoshina, S. et al. Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. *J. Biol. Chem.* **288**, 30161–30171 (2013).
49. Lu, X. J. & Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* **3**, 1213–1227 (2008).

Acknowledgements We thank the Electron Microscopy Laboratory of Peking University (cryo-EM platform) and the Tsinghua University Branch of the China National Center for Protein Sciences (Beijing) for the data collection of the ORC–DNA and apoORC samples, respectively. The computation was supported by High-performance Computing Platform of Peking University. This work was supported by the Ministry of Science and Technology of China (2016YFA0500700 to N.G.), the National Natural Science Foundation of China (NSFC) (31761163004, 31725007 and 31630087 to N.G.; 31700655 to N.L.), the Research Grants Council (RGC) of Hong Kong (GRF16138716 to B.K.T.; GRF16104115, GRF16143016 and GRF16104617 to Y.L.Z. and B.K.T.), NSFC/RGC Joint Research Scheme (N_HKUST614/17 to N.G., B.K.T., Y.L.Z. and N.L.), and the China Postdoctoral Science Foundation (2017M610013 to N.L.). N.L. is supported by Young Elite Scientists Sponsorship Program by CAST and a postdoctoral fellowship from the Peking-Tsinghua Centre for Life Sciences.

Reviewer information *Nature* thanks C. Fox and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.L.Z., N.G. and B.K.T. conceived the study; W.H.L., Y.L.Z. and Y.Q.Z. purified ORC; J.C., N.L., E.C., W.H.L. and Y.L.Z. prepared cryo grids; N.L. and J.C. collected data; N.L., J.C., Y.L.Z. and N.G. processed data; and N.L., W.H.L., Y.L.Z., N.G. and B.K.T. prepared the manuscript. N.L., W.H.L., Y.L.Z. and J.C. contributed equally to the study.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0293-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0293-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Y.L.Z. or N.G. or B.K.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

ORC purification. Orc1–Orc6 was purified with a yeast strain ySD-ORC (a gift from J. Diffley) as previously described⁵⁰ with the following modifications. In brief, 9 l of yeast cells were cultured in YPA medium containing 2% (w/v) raffinose at 30 °C until OD₆₀₀ reached 3.0. Alpha factor was then added to a final concentration of 100 ng ml⁻¹ to synchronize cells at G1 phase before overexpression induction of Orc1–Orc6 by addition of 2% (w/v) D-galactose for 3 h. Cells were collected and the pellet was washed with ice-cold water twice, and resuspended in 50 ml lysis buffer (25 mM HEPES-KOH pH 7.6, 0.5 M KCl, 10% glycerol, 0.05% (v/v) NP-40, 1× PI cocktail (Roche) and 1 mM PMSF). The cell suspension was frozen drop-wise in liquid nitrogen. Popcorns of the frozen cells were then crushed using a freezer mill (SPEX CertiPrep 6850 Freezer/Mill) with a setting of 15 cps, 2 min on, 2 min off, for 6 cycles. The powder was thawed on ice with addition of an equal volume of lysis buffer. The suspension was centrifuged twice for clarification at 38,900 for 20 min at 4 °C with a R20A2 rotor. The supernatant was incubated with 5 ml of 50% slurry of calmodulin affinity resin (Agilent Technologies) in the presence of 2 mM CaCl₂ for 4 h at 4 °C. After removal of flow-through and extensive washing of the resin with 50 column volumes of ice-cold washing buffer (25 mM HEPES-KOH pH 7.6, 0.3 M KCl, 10% glycerol, 2 mM CaCl₂ and 0.05% (v/v) NP-40), elution was performed by incubating the resins with 4 ml of elution buffer (25 mM HEPES-KOH pH 7.6, 0.3 M KCl, 10% glycerol, 1 mM EDTA, 2 mM EGTA and 0.05% (v/v) NP-40) for 30 min at 4 °C. The eluent was concentrated and further fractionated with size-exclusion chromatography (SEC) with Superose 6 10/300 GL column (GE Healthcare) in SEC buffer (25 mM HEPES-KOH pH 7.6, 0.15 M KCl, 10% glycerol, 2 mM β-mercaptoethanol, 0.05% (v/v) NP-40). Peak fractions containing the target complex were pooled and frozen in liquid nitrogen for storage.

Preparation of the ORC-DNA sample. DNA oligonucleotides used in this study (36 bp forward: 5'-TGGTTTTATATGTTATGATTGTTATT-3' and 36 bp reverse: 5'-AAAATAAACATACATAACAAAACATATAAAACCA-3', 72 bp forward: 5'-TGGTTTTATATGTTTGTATGTATTGTTATTTCCTTAAATTTAGGATATGAAAACAAGAATTATC-3' and 72 bp reverse: 5'-GATAATTCTTGTTCATATCCTAAATTAAAGGAAATAAACATACATAACAAAACATATAAAACCA-3') were first dissolved in a buffer containing 10 mM HEPES-KOH (pH 7.5), 1 mM EDTA and 50 mM NaCl to a final concentration of 100 μM. Equal volumes of the pairs of dissolved DNA oligonucleotides were mixed and incubated at 94 °C for 5 min. The reaction was left in the heat block to cool naturally to room temperature for DNA annealing. To assemble ORC onto the annealed origin DNA, 370 μl of 1.6 μM ORC was first mixed with an equal volumes of a buffer containing 25 mM HEPES-KOH pH 7.6, 0.1 M potassium acetate, 8 mM magnesium acetate, 0.02% (v/v) NP-40 and 2 mM ATPγS, and incubated on ice for 30 min. The purified ORC was unstable and prone to dissociate into sub-complexes (Extended Data Fig. 1b). To preserve the integrity of the assembled ORC-DNA, a mild fixation was applied to the complexes before cryo-EM grid preparation. Specifically, after a further incubation for 30 min at room temperature, the mixture was concentrated to 100 μl and then applied on the top of a 2-ml glycerol (10–30%) gradient containing glutaraldehyde (0–0.025%) in a buffer (25 mM HEPES-KOH pH 7.6, 0.1 M potassium acetate, 8 mM magnesium acetate, 0.5 mM β-mercaptoethanol) for GraFix⁵¹. The gradient was centrifuged in a Beckman TLS55 rotor (Beckman Optima TLX ultracentrifuge) for 13.5 h at a speed of 82,000g at 4 °C. Fractions were collected and the crosslinking reaction was quenched by addition of ice-cold Tris-HCl (pH 8.0) to a final concentration of 40 mM. For apoORC sample, the ORC protein was prepared in the same way but without incubation with ATPγS and dsDNA oligonucleotides. The GraFix treatment dramatically stabilized the ORC-DNA sample on cryo-EM grids (Extended Data Fig. 1d).

Electron microscopy. Fractions containing intact ORC-DNA (or ORC) complexes (Extended Data Fig. 1a) were pooled. Ultrafiltration for removal of glycerol and buffer exchange (25 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 5 mM magnesium acetate, 0.5 mM β-mercaptoethanol) was performed with a centrifugal filter (Amicon Ultra-0.5ml 50 k) at 6,000g at 4 °C. For negative staining, 4-μl aliquots of samples in serial dilutions were applied onto copper grids, washed and stained with 2% uranyl acetate. The negative-stained grids were examined using an FEI Tecnai T12 electron microscope operated at 120 kV, equipped with a 4 k × 4 k CCD camera (Gatan). The relative concentration of the samples used for cryo-grid preparation (about 10 times higher) were estimated from negative-staining electron microscopy. Aliquots (4 μl) of samples were applied onto the glow-discharged holey-carbon gold grids (Quantifoil, R1.2/1.3, 300 mesh) and blotted using an FEI Vitrobot Mark IV. Cryo-grids were screened using an FEI Talos Arctica operated at 200 kV with an FEI Ceta camera. Good cryo-grids were recovered and stored in liquid nitrogen. For the ORC-DNA sample, data were collected using a Titan Krios microscope (FEI) operated at 300 kV and images were collected using Leginon⁵² with a nominal magnification of 130,000× and a defocus range of 1.5–2.5 μm. The

images were recorded using a K2 summit camera equipped with GIF Quantum energy filter (Gatan) in the super-resolution counting and movie mode, with a dose rate of 4.4 e⁻ s⁻¹ Å⁻² and a total exposure time of 12 s. Each movie stack contains 40 frames with a calibrated pixel size of 0.526 Å at object scale (super-resolution). For the sample of apoORC, data acquisition was performed on a Titan Krios operated at 300 kV with defocus ranging from 1.5 to 2.5 μm. Images were recorded using a K2 summit camera (Gatan) in a super-resolution counting mode at a nominal magnification of 22,500× corresponding to a calibrated pixel size of 0.66 Å at object scale. Data were acquired semi-automatically using UCSF-Image4⁵³ in a movie mode with a dose rate of 10.2 e⁻ s⁻¹ Å⁻² and a total exposure time of 8 s, rendering a movie stack of 32 frames for each micrograph.

Image processing. A total of 1,765 movie stacks were acquired for the ORC-DNA (36 bp) sample (Extended Data Fig. 2a). Drift-correction, electron-dose weighting and two-fold binning were applied to the movie stacks using MotionCor2⁵⁴, which generate summed images with or without dose weighting. Program of CTFFIND4⁵⁵ was used to evaluate the parameters of contrast transfer function (CTF) of each micrograph based on images without dose weighting. Summed images and CTF power spectra were screened manually and high quality images were kept for further processing. Around 2,000 particles were manually picked and subjected to two-dimensional (2D) classification using RELION2.0⁵⁶ to provide templates for large-scale particle auto-picking. With the templates, 478,000 particles were auto-picked from images without dose weighting using RELION2.0. After two rounds of 2D classification, 464,000 particles were kept for further processing (Extended Data Fig. 2a). The initial three-dimensional (3D) model was calculated using CryoSPARC⁵⁷. To improve the performance of 3D classification, one round of 3D refinement was first performed on all particles selected after 2D classification using RELION2.0. Particles were re-centred according to the parameters from the last iteration of 3D refinement and re-extracted from dose-weighted images (particles that are too close to or exceed image edges were discarded). The first round of 3D classification was applied to exclude particles with obvious compositional or conformational heterogeneity. Around 42% good particles (196,000) were selected for further fine 3D classification with different parameters tested to separate particles corresponding to different conformational states as completely as possible (Extended Data Fig. 2a). Three different conformational states (I, II and III) were finally separated. Among these three, state I contained approximately 52% of particles and displayed more high-resolution features; states II and III each had ~12% of particles. States II and III were further refined to final resolutions of 4.7 Å and 5.6 Å, respectively. In the map of state II, Orc2-WHD moves towards the WHD tier of the ORC and the gap between Orc1 and Orc2 AAA+ domain is slightly smaller than State I (Extended Data Fig. 6b, d). While in the map of state III, Orc2-WHD is in the same position as that of state I but very unstable and the Orc1/Orc2 gap becomes larger compared to state I (Extended Data Fig. 6c, e). For state I, 3D refinement with a global mask applied improved the resolution to 3.8 Å. Application of particle-level local defocus evaluated using Gctf⁵⁸ based on images without dose weighting further improved the resolution to 3.6 Å. The resolution was estimated by gold-standard FSC 0.143 criteria, after correction of the mask effect. The map was sharpened by an auto-evaluated B-factors using RELION2.0. The local resolution map was created using ResMap⁵⁹ and displayed using UCSF Chimera⁶⁰.

For the ORC-DNA (72 bp) dataset, 3,603 micrographs were similarly processed (Extended Data Fig. 1d–f). A total of 427,000 particles were selected after 2D classification. After 3D classification, three classes (state I) were combined (approximately 164,000 particles) and subjected to high-resolution refinement, resulting in a final density map at a resolution of 3.0 Å. A major structural difference in other classes (states II and IV) from the ORC-DNA (72 bp) dataset is the extent of DNA curving (Extended Data Fig. 6f, g).

Micrographs (1,041 in total) of apoORC complexes were also similarly processed (Extended Data Fig. 2b). In total, 262,000 particles were auto-picked, and 189,000 particles were selected after 2D classification. After two rounds of 3D classification, 45,000 good particles were selected for final refinement, resulting in a final density map at a resolution of 8.2 Å. The overall subunit arrangement of this apoORC is in general the same as that of the origin-bound complex, but ORC alone displayed more tendency to dissociate during sample preparation even when GraFix was applied, indicating that DNA binding stabilizes the whole complex.

Model building. Modelling was first performed using the map of state I from the 36-bp dataset. To facilitate the modelling, multiple versions of density maps were obtained by applying additional rounds of mask-based refinement using RELION2.0. For example, to improve the resolution of the core region, a mask excluding Orc6, the Orc3 insertion domain and unstable portions of dsDNA was applied. These additional procedures have improved the quality of local densities, although the reported resolutions had no marked gain (Extended Data Fig. 2a, c).

The initial template used for the modelling was from the OCCM structure (PDB code 5UDB)³⁰. Each domain of the Orc1–Orc5 subunits was manually fitted into the density map in Chimera, followed by manual rebuilding using Coot⁶¹.

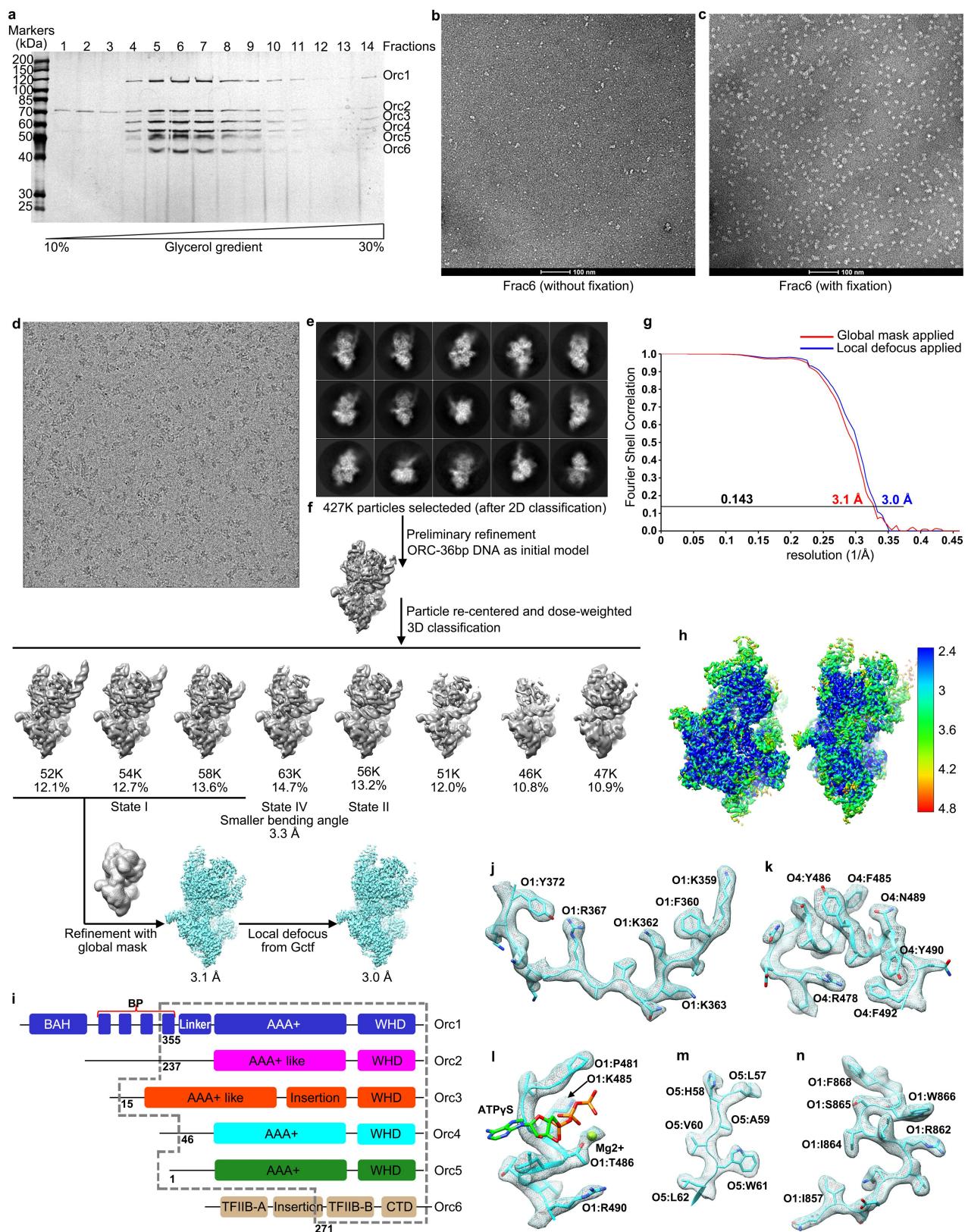
Linkers, flexible extensions and TFIIB-B domain of Orc6 were modelled de novo. Secondary structure prediction of the Orc1–Orc6 subunits was performed using PSIPRED⁶². For modelling the origin DNA, a 36-bp ideal B-form dsDNA (with ARS305) was generated, fitted in the density map and manually adjusted using Coot. The model was refined against the 3.6-Å density map using Phenix.real_space_refine⁶³ with secondary structure restraints, geometry restraints and DNA-specific restraints applied. Chains of the model were then extended with the 3.0-Å map from the 72-bp dataset, refined and optimized similarly. The final models were evaluated by MolProbity⁶⁴, and statistics are presented in Extended Data Table 1.

For the 8.2 Å map of the apoORC, each domain of the Orc1–Orc6 subunits from the atom model of the ORC–DNA structure was fitted into the density map in Chimera, generating a pseudo-atomic model for the apoORC complex. Chimera and Pymol (<http://pymol.org>) were used for figure preparation.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Cryo-EM maps of the apoORC, ORC–36-bp and ORC–72-bp complexes have been deposited in the Electron Microscopy Data Bank (EMDB) with accession codes EMD-6943, EMD-6942 and EMD-6941, respectively. Atomic coordinates of the ORC–72-bp complex have been deposited in the Protein Data Bank (PDB) with accession code 5ZR1. All other data are available from the corresponding author upon reasonable request.

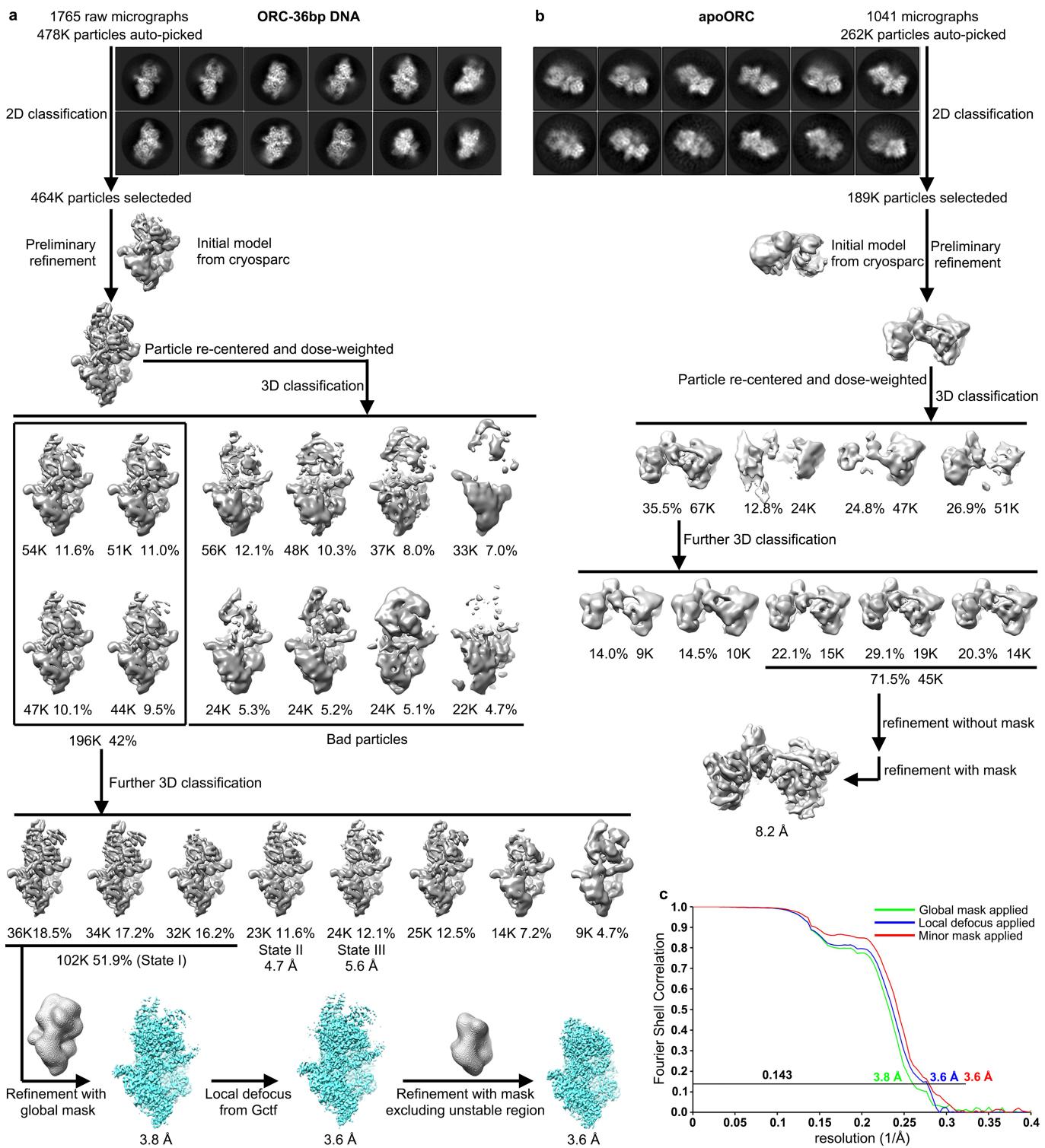
50. Frigola, J., Remus, D., Mehanna, A. & Diffley, J. F. ATPase-dependent quality control of DNA replication origin licensing. *Nature* **495**, 339–343 (2013).
51. Kastner, B. et al. GraFix: sample preparation for single-particle electron cryomicroscopy. *Nat. Methods* **5**, 53–55 (2008).
52. Súloway, C. et al. Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151**, 41–60 (2005).
53. Li, X., Zheng, S., Agard, D. A. & Cheng, Y. Asynchronous data acquisition and on-the-fly analysis of dose fractionated cryoEM images by UCSFImage. *J. Struct. Biol.* **192**, 174–178 (2015).
54. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
55. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
56. Kimanis, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
57. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
58. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
59. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
60. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
61. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
62. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349–57 (2013).
63. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
64. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
65. Liu, S. et al. Structural analysis of human Orc6 protein reveals a homology with transcription factor TFIIB. *Proc. Natl. Acad. Sci. USA* **108**, 7373–7378 (2011).
66. Nikolov, D. B. et al. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* **377**, 119–128 (1995).



Extended Data Fig. 1 | See next page for caption.

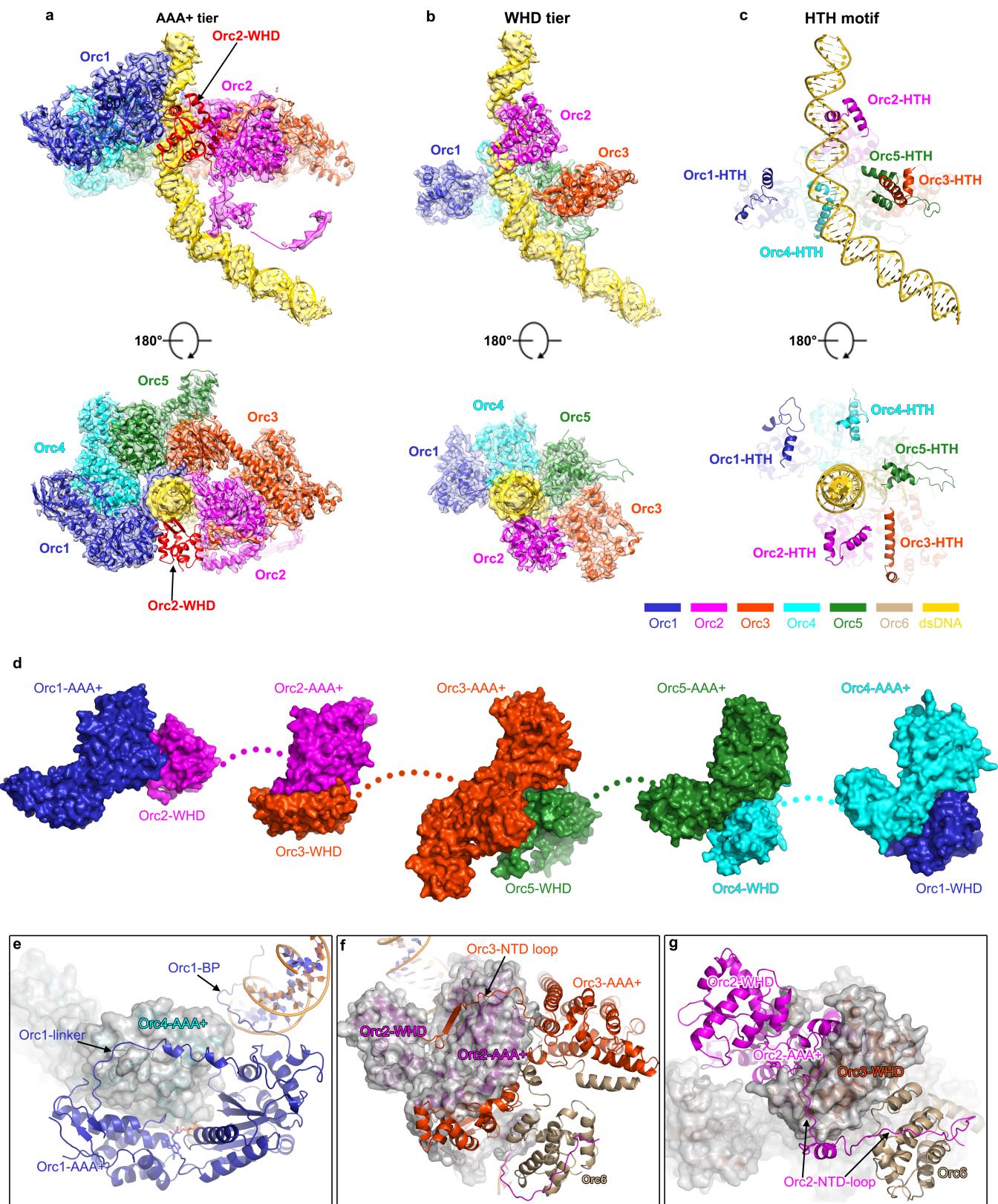
Extended Data Fig. 1 | Sample preparation and image processing of the ORC–DNA complex. **a**, SDS–PAGE analysis of the glycerol gradient fractions. ORC–DNA complexes (no fixation) were subjected to 10–30% glycerol gradient centrifugation. Fractions were collected and resolved on SDS–PAGE. Peak fractions (5–7) containing intact ORC complexes were processed for further electron microscopy analysis. Experiments were repeated multiple times ($n > 10$), and similar results were obtained. **b**, Negative-staining electron microscopy of the ORC–DNA complex. Samples from fraction 6 were subjected to negative staining. A severe dissociation of complexes was observed. **c**, Negative-staining electron microscopy of the ORC–DNA complex prepared with GraFix method using a gradient of glutaraldehyde (0–0.025%). **d**, A representative raw

cryo-EM image of the ORC–DNA (72 bp) complex. **e**, 2D class averages of the ORC–DNA (72 bp) particles. **f**, Workflow of image processing of the ORC–DNA (72 bp) particles. The processing includes rounds of 2D classification, 3D classification, structural refinement and masked-based refinement procedures. **g**, FSC curves of the final density map of the ORC–DNA (72 bp) complex. **h**, The local resolution map of the final density map. **i**, Schematic domain organization of Orc1–Orc6 subunits. Regions that were built in the final atomic model were boxed in dashed grey lines. **j–n**, Local density of representative regions of the final cryo-EM density map, for Orc1-BP (**j**), Orc4-IH (**k**), the Orc1 ATP-binding pocket (**l**) and two other regions (**m**, **n**). For clarity, density of ATP γ S is omitted in **j** to highlight the Walker A motif of Orc1.



Extended Data Fig. 2 | Workflow of the image processing of the ORC-DNA (36 bp) and apoORC particles. **a**, Image processing workflow of the ORC-DNA (36 bp) particles. Processing includes rounds of 2D classification, 3D classification, structural refinement and masked-based

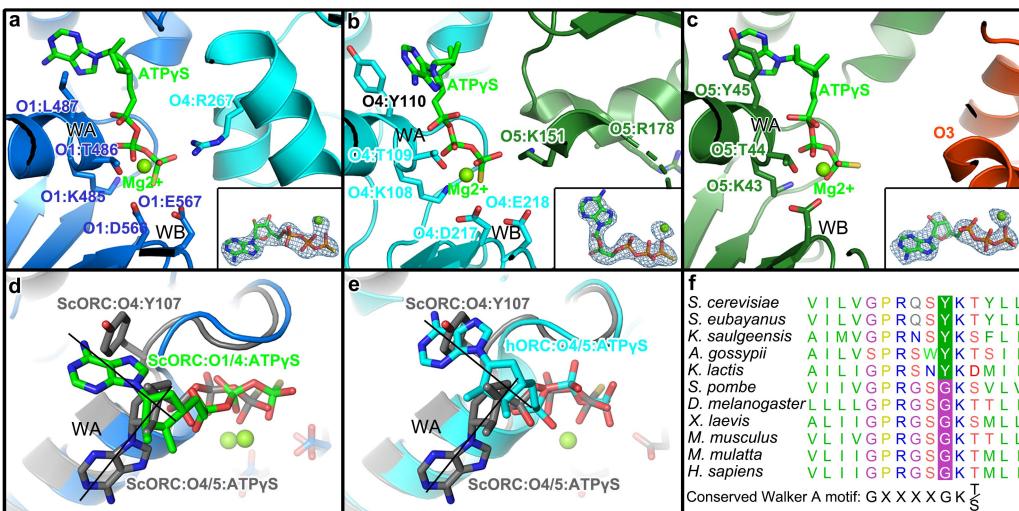
refinement procedures. **b**, Image processing workflow of the apoORC particles. **c**, FSC curves of the density maps of the ORC-DNA (36 bp) complex.



Extended Data Fig. 3 | See next page for caption.

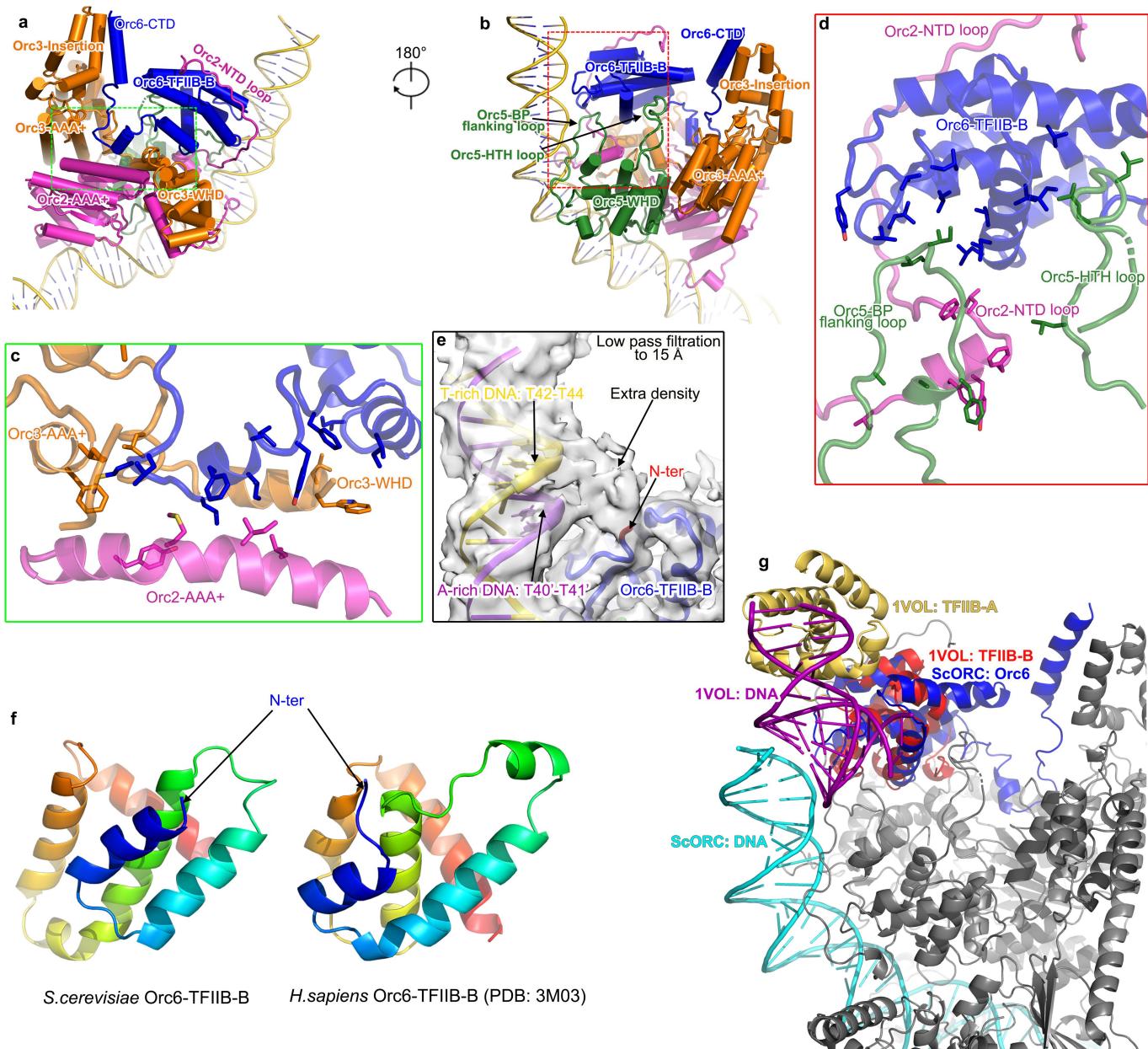
Extended Data Fig. 3 | Organization of AAA+ and WHD domains around the origin DNA. **a**, Organization of AAA+ domains of Orc1–Orc5 subunits around origin DNA. Cryo-EM maps of the AAA+ domains and DNA are shown in solid surface representation and colour-coded. The WHD of Orc2, which blocks the gap between the AAA+ domains of Orc1 and Orc2 is shown in cartoon representation. **b**, Same as in **a**, but for the WHDs of the Orc1–Orc5 subunits. **c**, Distribution of the HTH motifs of WHDs around the origin DNA. The WHDs of the Orc1–Orc5 subunits are shown in cartoon representation with the HTH motifs highlighted. **d**, Domain swapping between the AAA+ and WHD tiers. As shown, Orc2-WHD is in a different position from the rest. **e**, A flexible

linker (residues 375–436) upstream of Orc1 AAA+ domains extends on the surface of Orc4 AAA+ domain (surface representation), with the further upstream basic patch sequences inserted into the minor groove of the ACS. The bound ATP γ S is shown in stick model (orange). **f**, The very N-terminal extension (residues 15–50) of Orc3 wraps around the AAA+ domain of Orc2 (surface representation) and ends in the interface between the Orc2-WHD and Orc2-AAA+ domain. **g**, A very long N-terminal linker (NTD loop) upstream AAA+ domain of Orc2 extends on the surface of Orc3-WHD and TFIIB-B domain of Orc6. Note that the linker of Orc2 wrapping around Orc6 is traceable in the cryo-EM density map but the model could not be built at atomic level.



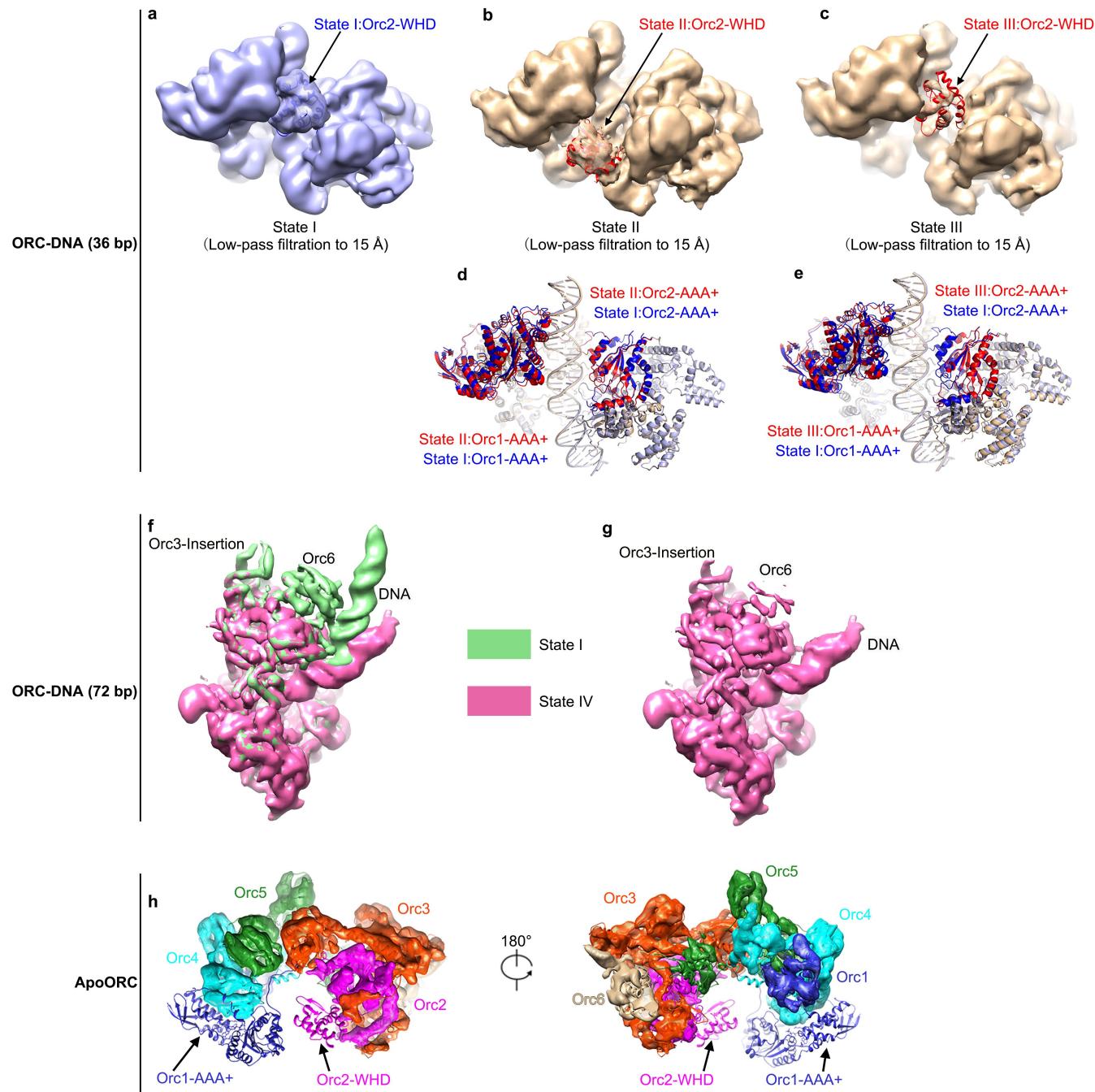
Extended Data Fig. 4 | Configuration of the three ATPase centres in the ORC-DNA complex. **a**, Zoomed-in view of the ATPase centre formed between Orc1 (O1) and Orc4 (O4). Orc1, Orc4 and ATP γ S-Mg $^{2+}$ are coloured blue, cyan and green, respectively. The Walker A and B motifs (WA and WB) of Orc1 and the arginine finger of Orc4 (R267) are highlighted in stick models. Inset, the stick model of ATP γ S-Mg $^{2+}$ superimposed with the cryo-EM density. **b**, Zoomed-in view of the ATPase centre formed between Orc4 and Orc5 (O5). Orc4, Orc5 and ATP γ S-Mg $^{2+}$ are coloured cyan, dark green and green, respectively. The Walker A and B motifs (WA and WB) of Orc4 and the equivalent arginine finger of Orc5 (R178) are highlighted in stick models. K151 of Orc5 within 4 Å distance from the γ -phosphate is shown. Inset, the stick model of ATP γ S-Mg $^{2+}$ superimposed with the cryo-EM density. **c**, Zoomed-in view of the ATPase centre formed between Orc5 and Orc3 (O3). Orc5, Orc3 and ATP γ S-Mg $^{2+}$

are coloured dark green, orange and green, respectively. The Walker A and B motifs of Orc5 are highlighted in stick models. Inset, the stick model of ATP γ S-Mg $^{2+}$ superimposed with the cryo-EM density. **d**, Comparison between the ATPase centres of O1:O4 and O4:O5, highlighting the flip of the base moiety of the bound ATP γ S within the O4:O5 centre. The flip is forced by the replacement of a conserved glycine by a bulky tyrosine residue (Y107) of the Walker A motif of Orc4. The Walker A motifs were used as reference in the alignment. **e**, Comparison between the ATPase centres of the yeast O4:O5 and human O4:O5 (PDB code 5UJ7)³², highlighting the flip of the base moiety of the bound ATP γ S within the yeast O4:O5 centre. The Walker A motifs were used as reference in the alignment. **f**, Sequence alignment of the Walker-A motif of Orc4 from different species.


Extended Data Fig. 5 | Orc6 interacts with Orc2, Orc3 and Orc5.

a, b, Overview of the interactions between Orc3, Orc2, Orc5 and Orc6. **c, d**, Zoomed-in views of the boxed regions in **a** and **b** to highlight their relatively hydrophobic interfaces. Selected hydrophobic residues at the interface are displayed in stick model. A short helix in the linker between Orc6-CTD and Orc6-TFIIB-B packs with two helices from Orc2-AAA+ and Orc3-WHD (c). A long N-terminal linker of Orc2 (upstream the AAA+ domain) wraps the TFIIB-B domain of Orc6. Note that the linker of Orc2 (Orc2-NTD loop) is traceable in the cryo-EM density map but the model could not be built at atomic level. **e**, Low-pass filtered map of the ORC-DNA complex, highlighting the interactions (indicated by the presence of extra density) between the linker sequence of Orc6

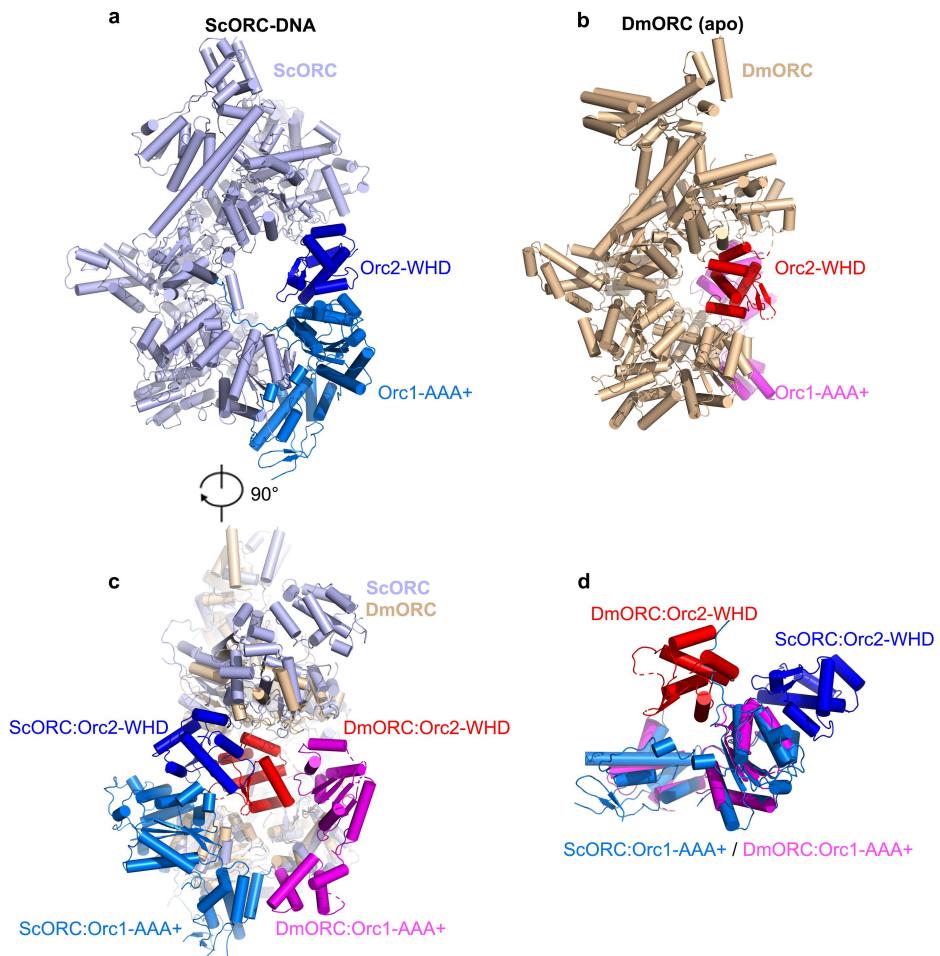
(between TFIIB domains A and B) and DNA. The N-terminal (N-ter) end of the model built for Orc6 in our map is S217. **f**, Comparison between the yeast Orc6-TFIIB-B and human ORC6-TFIIB-B. The structure of human ORC6 is from a crystallography study (PDB code 3M03)⁶⁵. The overall structure of the yeast ORC6-TFIIB-B is quite similar to its human counterpart. **g**, Superimposition of the structure of TFIIB-DNA onto the ORC-DNA complex. The crystal structure of a human TFIIB-TBP-DNA (PDB code 1VOL)⁶⁶ was aligned using ORC6-TFIIB-B as reference. As shown, ORC6-TFIIB-B has not established extensive interactions with DNA. It is possible that further conformational change in Orc6 is required to form extensive interactions with DNA as the TFIIB does, probably at a later stage of replication licensing.



Extended Data Fig. 6 | Flexibility of the ORC complexes.

a–c, Comparison of states I, II and III of the ORC–DNA complex (36 bp). Density maps of the three states are displayed in surface representation and in the Orc1–Orc2 side-view. The model of Orc2-WHD is highlighted in red cartoon. As shown, Orc2-WHD occupies different positions in the three maps. In the map of state II, the density of Orc2-WHD is relatively weak and it takes a position similar to that of the OCCM structure³⁰. In the map of state III, Orc2-WHD is in a similar position as in state I, but its density is highly fragmented. Together, these indicate a floppy nature of the Orc2-WHD. **d**, Superimposition of the models of states I and II. The atomic model of state II was derived by flexible fitting of the state I model into the density map of state II. The alignment was done using Orc2 and Orc3. Compared with state I, the opening of the gap in the structure of state II is narrower. For clarity, the WHDs of Orc2 in the two states are omitted. **e**, Superimposition of the models of states I and III. The atomic

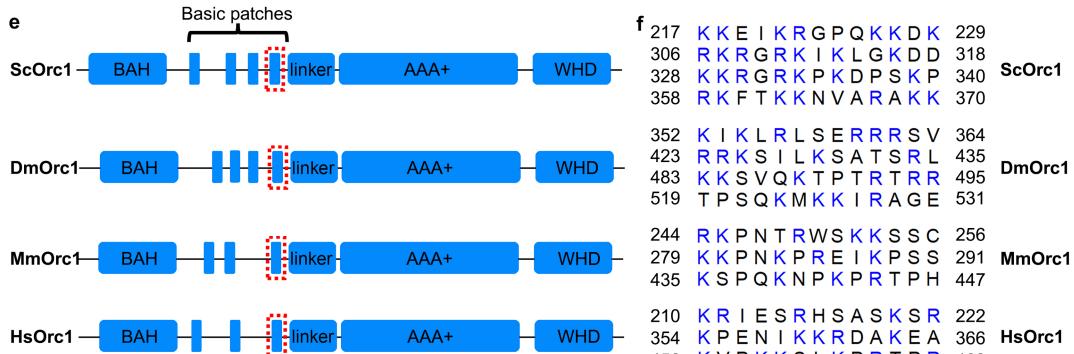
model of state III was derived by flexible fitting of the state I model into the density map of state III. The alignment was done using Orc2 and Orc3. Compared with state I, the opening of the gap in the structure of state III is slightly larger. For clarity, the WHDs of Orc2 in the two states are omitted. **f, g**, Comparison of the density maps of states I and IV from the ORC–DNA (72 bp) dataset. A major difference between the two maps is the bending angle of the DNA. The extent of DNA bending correlates with the stability of Orc6 and Orc3 (Insertion domain of the AAA+ module). **h**, Top (left) and bottom (right) views of the cryo-EM map of the apoORC complex with the atomic model superimposed, which was derived by flexible fitting of the ORC–DNA model into the density map. ORC subunits are colour-coded. The AAA+ domain of Orc1 and the WHD of Orc2 are highly flexible, resulting in a large opening between Orc1 and Orc2, as indicated by the reduced EM densities of the corresponding regions.



Extended Data Fig. 7 | Structural comparison between the *S. cerevisiae* ORC-DNA and the *Drosophila* apoORC complexes. **a, b**, Side-by-side comparison of the yeast ORC-DNA and the *Drosophila* apoORC (PDB code 4XGC)³¹ complexes. **a**, The yeast ORC-DNA structure is shown in cartoon representation, with Orc1-AAA+ and Orc2-WHD highlighted in marine and blue, respectively. **b**, The *Drosophila* apoORC structure is shown in cartoon representation, with Orc1-AAA+ and Orc2-WHD highlighted in magenta and red, respectively. **c**, Superimposition of the

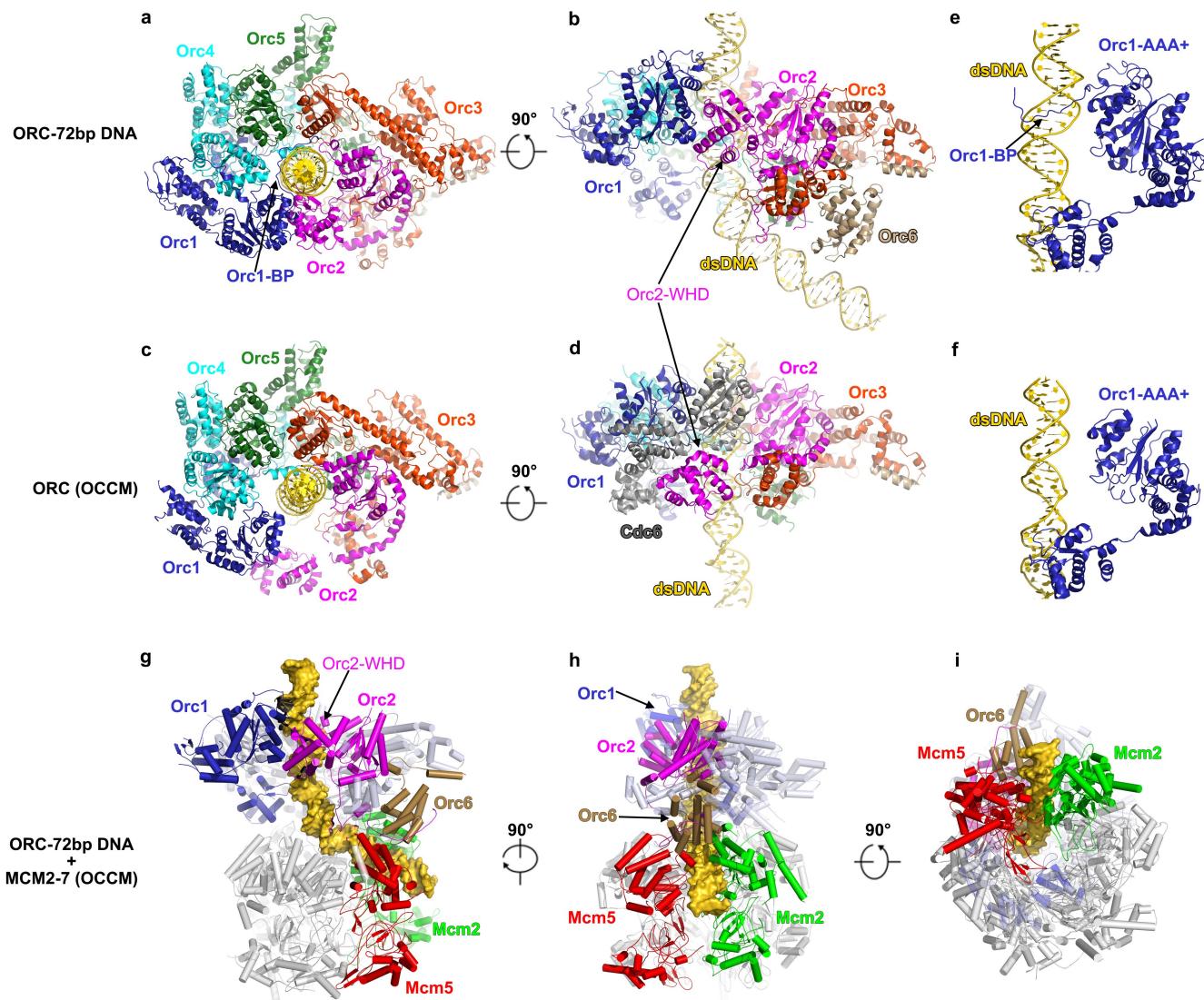
yeast ORC-DNA and the *Drosophila* apoORC structures using Orc3–Orc5 as reference. Note that the positions and orientations of Orc1-AAA+ and Orc2-WHD are markedly different in the two structures. **d**, Superimposition of Orc1 from the structures of the yeast ORC-DNA and the *Drosophila* apoORC complexes using Orc1-AAA+ as a reference. Note that the Orc2-WHDS in two structures are in totally different positions relative to Orc1-AAA+, highlighting the distinct interfaces between Orc1-AAA+ and Orc2-WHD in the two structures.

a Orc1-BP



Extended Data Fig. 8 | Multiple sequence alignment of Orc1-BP, Orc4-IH, Orc5-BP and Orc2-BP. **a-d**, Multi-sequence alignment of Orc1 N-terminal patches (**a**), Orc4 insertion helixes (**b**), Orc5 WHD basic patches (**c**) and Orc2 N-terminal basic patches (**d**) from various species as indicated. **e**, Multiple basic patches found between the BAH and AAA+ domains of Orc1 from yeast to human. The criteria for basic patches are

a stretch of 10 to 14 amino acids flanked by either lysine or arginine with at least three basic (K or R) residues in between and a pair of them are spaced 3–4 residues apart as found in Orc1 (R367 and K362). **f**, Sequence information of the Orc1 basic patches in **d** from various species are listed as indicated.



Extended Data Fig. 9 | Structural comparison between the yeast ORC-DNA and OCCM complexes. **a, b**, AAA+ (a) and side (b) views of the ORC-DNA complex. ORC subunits and DNA are shown in cartoon representation and colour-coded. **c, d**, AAA+ (c) and side (d) views of the ORC in the context of the OCCM complex. ORC subunits and DNA are shown in cartoon representation and colour-coded. The OCCM structure (PDB code 5UDB) is from previous cryo-EM work³⁰. Compared with the OCCM structure, ORC subunits of Orc1 and Orc2 in the structure of ORC-DNA are more compact around the DNA. Cdc6 (grey) is included

in the side view. **e**, Relative orientation of the origin DNA with Orc1 in the ORC-DNA complex. Orc1-BP is inserted into the minor groove of ACS DNA. **f**, Same as in **e**, but for the DNA and Orc1 in OCCM complex. The distance between the AAA+ domain and DNA is considerably larger, resulting in the loss of DNA contact. **g–i**, Superimposition of the ORC-DNA (72 bp) and OCCM (PDB code 5UDB)³⁰ structures. For clarity, ORC subunits from the OCCM is not shown. The Mcm2–Mcm7 subunits from the OCCM are shown in grey. Only Mcm2 and Mcm5 are labelled and colour-coded as indicated.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	ORC-72bp DNA (EMDB-6941) (PDB 5ZR1)	ORC-36bp DNA (EMDB-6942)	apoORC (EMDB-6943)
Data collection and processing			
Magnification	130,000	130,000	22,500
Voltage (kV)	300	300	300
Electron exposure (e ⁻ /Å ²)	52.3/dose weighting	52.3/dose weighting	50/dose weighting
Defocus range (μm)	1.3-2.3	1.3-2.3	1.5-2.5
Pixel size (Å)	1.052	1.052	1.32
Symmetry imposed	C1	C1	C1
Initial particle images (no.)	427K	464K	189K
Final particle images (no.)	164K	102K	45K
Map resolution (Å)	3.0	3.6	8.2
FSC threshold	0.143	0.143	0.143
Map resolution range (Å)	2.4-4.9	3.4-4.9	5.8-10.8
Refinement			
Initial model used (PDB code)	Built based on atom model of OCCM (5UDB) or <i>de novo</i>	Built based on atom model of OCCM (5UDB) or <i>de novo</i>	
Map sharpening <i>B</i> factor (Å ²)	-94	-142	-167
Model composition			
Non-hydrogen atoms	22260	21850	
Protein and DNA residues	2479	2448	
Ligands (ATPγS and Mg ²⁺)	6	6	
R.m.s. deviations			
Bond lengths (Å)	0.0069	0.0098	
Bond angles (°)	1.29	1.50	
Validation			
MolProbity score	1.49	1.79	
Clashscore	4.10	4.46	
Poor rotamers (%)	0.65	1.49	
Ramachandran plot			
Favored (%)	95.68	92.65	
Allowed (%)	4.32	7.36	
Disallowed (%)	0.00	0.29	

Non-gravitational acceleration in the trajectory of 1I/2017 U1 ('Oumuamua)

Marco Micheli^{1,2*}, Davide Farnocchia³, Karen J. Meech⁴, Marc W. Buie⁵, Olivier R. Hainaut⁶, Dina Prialnik⁷, Norbert Schorghofer⁸, Harold A. Weaver⁹, Paul W. Chodas³, Jan T. Kleyna⁴, Robert Weryk⁴, Richard J. Wainscoat⁴, Harald Ebeling⁴, Jacqueline V. Keane⁴, Kenneth C. Chambers⁴, Detlef Koschny^{1,10,11} & Anastassios E. Petropoulos³

'Oumuamua (1I/2017 U1) is the first known object of interstellar origin to have entered the Solar System on an unbound and hyperbolic trajectory with respect to the Sun¹. Various physical observations collected during its visit to the Solar System showed that it has an unusually elongated shape and a tumbling rotation state^{1–4} and that the physical properties of its surface resemble those of cometary nuclei^{5,6}, even though it showed no evidence of cometary activity^{1,5,7}. The motion of all celestial bodies is governed mostly by gravity, but the trajectories of comets can also be affected by non-gravitational forces due to cometary outgassing⁸. Because non-gravitational accelerations are at least three to four orders of magnitude weaker than gravitational acceleration, the detection of any deviation from a purely gravity-driven trajectory requires high-quality astrometry over a long arc. As a result, non-gravitational effects have been measured on only a limited subset of the small-body population⁹. Here we report the detection, at 30σ significance, of non-gravitational acceleration in the motion of 'Oumuamua. We analyse imaging data from extensive observations by ground-based and orbiting facilities. This analysis rules out systematic biases and shows that all astrometric data can be described once a non-gravitational component representing a heliocentric radial acceleration proportional to r^{-2} or r^{-1} (where r is the heliocentric distance) is included in the model. After ruling out solar-radiation pressure, drag- and friction-like forces, interaction with solar wind for a highly magnetized object, and geometric effects originating from 'Oumuamua potentially being composed of several spatially separated bodies or having a pronounced offset between its photocentre and centre of mass, we find comet-like outgassing to be a physically viable explanation, provided that 'Oumuamua has thermal properties similar to comets.

The object now known as 1I/'Oumuamua was discovered on 2017 October 19 by the Pan-STARRS1 survey^{10,11}. Within a few days, additional observations collected with the European Space Agency (ESA) Optical Ground Station (OGS) telescope and at other observatories, together with pre-discovery data from Pan-STARRS1, allowed the determination of a preliminary orbit that was highly hyperbolic (eccentricity of 1.2). Such an orbit identified the object as originating from outside the Solar System¹ and approaching from the direction of the constellation Lyra, with an asymptotic inbound velocity of around 26 km s^{-1} .

This extreme eccentricity also led the Minor Planet Center to classify the object as a comet initially¹². However, this classification was later withdrawn when images obtained immediately after discovery using the Canada–France–Hawaii Telescope (CFHT) and, in the subsequent days, the European Southern Observatory (ESO) Very Large Telescope (VLT) and the Gemini South (GS) Telescope, both 8-metre-class facilities, found no sign of coma despite optimal seeing conditions (see Fig. 1 and

discussion in Methods). In addition, spectroscopic data obtained^{5,7} at around the same time showed no evidence of identifiable gas emission in the visible-wavelength region of the spectrum. Although the object has a surface reflectivity similar to comets^{1,5,7}, all other observational evidence available at the time suggested that 'Oumuamua was probably inactive and of asteroidal nature, despite predictions that cometary interstellar objects should be the easier to discover because they brighten more than asteroids^{1,13}.

In parallel with physical and compositional studies, our team continued to image 'Oumuamua to constrain its trajectory further through astrometric measurements. As 'Oumuamua faded, we obtained data with CFHT, VLT and the Hubble Space Telescope (HST; see Methods). A final set of images was obtained with HST in early 2018 for the purpose of extracting high-precision astrometry. The resulting dataset provides dense coverage from discovery to 2018 January 2, when the object became fainter than $V \approx 27$ at a heliocentric distance of 2.9 AU.

We analysed the full observational dataset, which includes 177 ground-based and 30 HST-based astrometric positions (for a total of 414 scalar measurements), applying the procedures and assumptions discussed in Methods. Our analysis shows that the observed orbital arc cannot be fitted in its entirety by a trajectory governed solely by gravitational forces due to the Sun, the eight planets, the Moon, Pluto, the 16 largest bodies in the asteroid main belt and relativistic effects¹⁴. As shown in Fig. 2a, the residuals in right ascension and declination of the best-fitting gravity-only trajectory are incompatible with the formal uncertainties: ten data points deviate by more than 5σ in at least one coordinate, and 25 are discrepant by more than 3σ . Furthermore, the offsets (as large as $22''$ for the 2017 October 14 Catalina observation) are not distributed randomly but show clear trends along the trajectory.

To improve the description of the trajectory of 'Oumuamua, we included a radial acceleration term $A_1 g(r)$ in the model⁸, where A_1 is a free fit parameter, r is the heliocentric distance and $g(r)$ is set to be proportional to r^{-2} , matching the decrease of solar flux with distance, with $g(1 \text{ AU}) = 1$. As shown in Fig. 2b, the addition of this term allows us to explain the data for $A_1 = (4.92 \pm 0.16) \times 10^{-6} \text{ m s}^{-2}$, which corresponds to a formal detection of non-gravitational acceleration with a significance of about 30σ . Additional analyses, discussed in greater detail in Methods, further support our finding that the non-gravitational acceleration is preferentially directed radially away from the Sun, and allow both the aforementioned r^{-2} dependency and a less steep r^{-1} law. By contrast, a constant acceleration independent of distance is strongly disfavoured, regardless of direction (radial, along the instantaneous velocity vector of 'Oumuamua or inertially fixed). Table 1 reports the χ^2 and reduced χ^2 (χ^2_ν) statistics for the astrometric fits of each of the models tested (see Methods for details). We used

¹ESA SSA-NEO Coordination Centre, Frascati, Italy. ²INAF—Osservatorio Astronomico di Roma, Monte Porzio Catone, Italy. ³Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. ⁴Institute for Astronomy, University of Hawai'i, Honolulu, HI, USA. ⁵Southwest Research Institute, Boulder, CO, USA. ⁶European Southern Observatory, Garching bei München, Germany. ⁷School of Geosciences, Sackler Faculty of Exact Sciences, Tel Aviv University, Ramat Aviv, Israel. ⁸Planetary Science Institute, Tucson, AZ, USA. ⁹The Johns Hopkins University Applied Physics Laboratory, Space Exploration Sector, Laurel, MD, USA. ¹⁰ESTEC, European Space Agency, Noordwijk, The Netherlands. ¹¹Chair of Astronautics, Technical University of Munich, Garching bei München, Germany. *e-mail: marco.micheli@esa.int

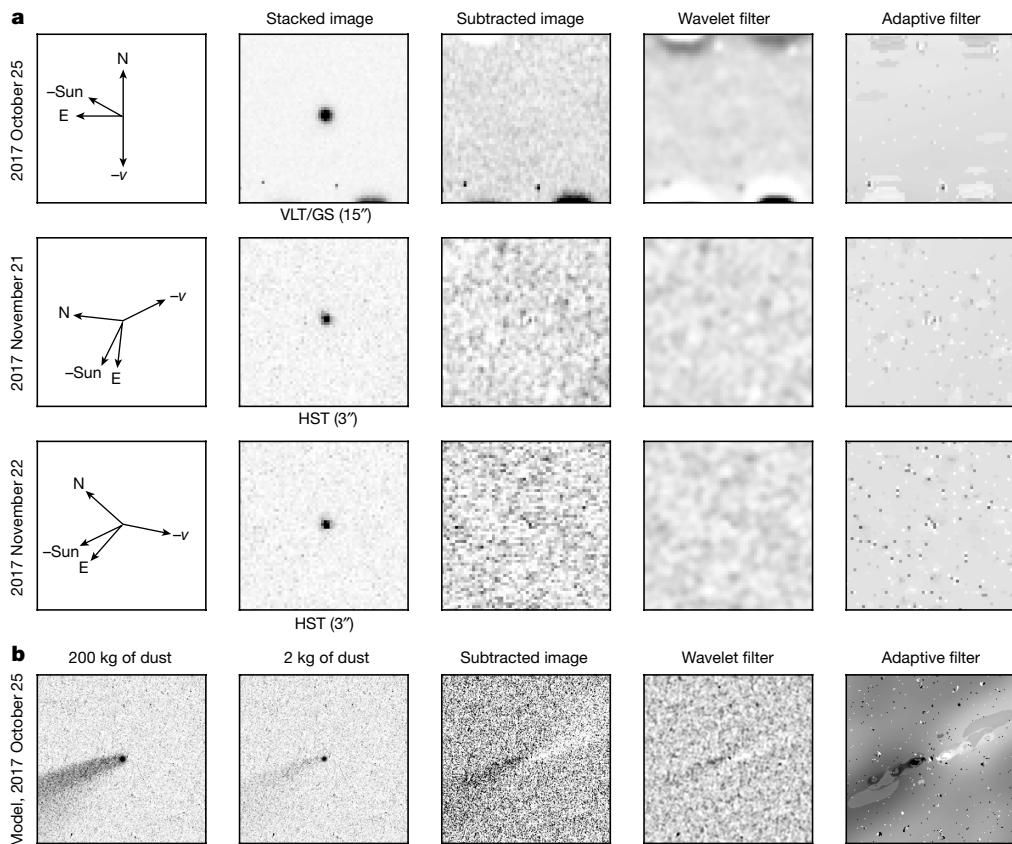


Fig. 1 | Deep stacked images for dust detection. **a**, For each date we show the image orientation (—Sun, anti-solar direction; $-\nu$, anti-motion direction), the stacked image (telescope and size of the image are listed below the image), a self-subtracted image (see Methods), and the image after application of a wavelet or adaptive filter to enhance low-surface-brightness features. No dust is visible. **b**, Images from a model

with an artificial cometary feature that matches the October geometry demonstrate the sensitivity of the image enhancement: a very strong dust feature is evident when 200 kg of dust is used in the point spread function (PSF) region (left-most panel); the other panels show the same feature scaled to 2 kg of dust in the PSF region (twice the observed ‘Oumuamua limit) and the image processed in the same manner as the real data.

conservative estimates for the measurement uncertainties that serve as data weights to mitigate the effect of systematic errors, for example, due to star catalogue biases, field-of-view distortions, clock errors or the absence of uncertainty information (for astrometry produced by others). As a result, the χ^2 and χ^2_ν values listed are lower than would be expected for purely Gaussian noise, and the correspondingly larger error bars that we derive more safely capture the actual uncertainties in the estimated parameters.

We performed a series of tests, also discussed in greater detail in Methods, which confirm that the non-gravitational signature is neither an artefact caused by some subset of the observations nor the result of overall systematic biases unaccounted for in the analysis. Even a substantial inflation of the assumed error bars in the astrometry, applied to reflect possible catalogue biases or uncorrected distortions, still results in a significant detection. In addition, the non-gravitational acceleration is clearly detected both in ground-based observations alone and

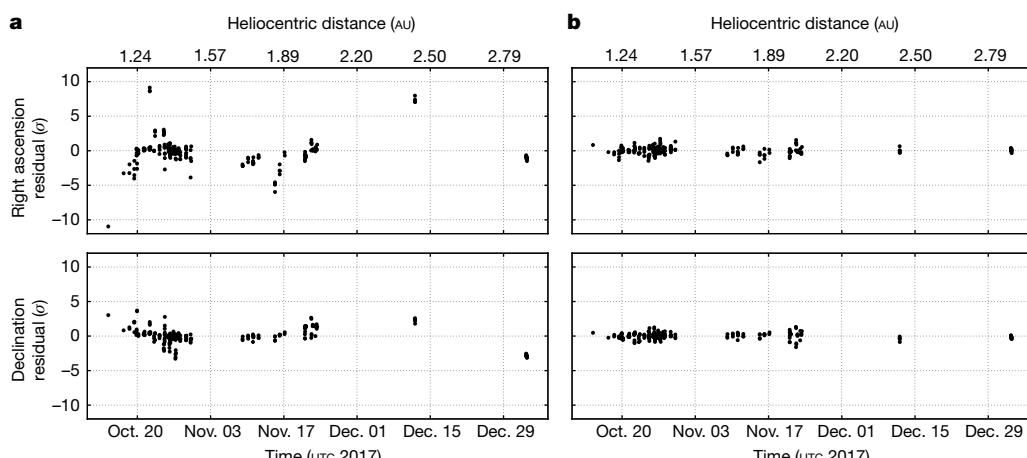


Fig. 2 | Astrometric residuals of ‘Oumuamua observations. **a, b**, Normalized residuals for right ascension and declination compared to a gravity-only solution (**a**) and a solution that includes a non-gravitational radial

acceleration of $A_1 r^{-2}$ (**b**). Because each residual is normalized to its formal uncertainty, each data point has a 1σ error bar (not shown) equal to 1 on this scale.

Table 1 | Fits for different non-gravitational models

Model	Number of parameters	χ^2	χ^2_ν
Gravity-only	6	1.031×10^3	2.53
(1) Impulsive change in velocity	10	117	0.29
(2) Pure radial acceleration, $A_1 g(r) \propto r^{-k}$, $k \in \{0, 1, 2, 3\}$	7	99, 80, 81, 98	0.24, 0.20, 0.20, 0.24
(3) RTN decomposition, $[A_1, A_2, A_3]g(r) \propto r^{-k}$, $k \in \{0, 1, 2, 3\}$	9	90, 80, 78, 87	0.22, 0.20, 0.19, 0.21
(4) ACN decomposition, $[A_A, A_C, A_N]g(r) \propto r^{-k}$, $k \in \{0, 1, 2, 3\}$	9	104, 85, 77, 83	0.26, 0.21, 0.19, 0.21
(5) Pure along-track acceleration, $A_{AG}(r) \propto r^{-k}$, $k \in \{0, 1, 2, 3\}$	7	1.031×10^3 , 1.025×10^3 , 1.002×10^3 , 963	2.53, 2.52, 2.46, 2.37
(6) Constant, inertially fixed acceleration vector	9	116	0.29
(7a) Pure radial acceleration, $A_1 g_{CO}(r)$	7	84	0.21
(7b) Pure radial acceleration, $A_1 g_{H2O}(r)$	7	111	0.27
(7c) RTN decomposition, $[A_1, A_2, A_3]g_{CO}(r)$	9	79	0.19
(7d) RTN decomposition, $[A_1, A_2, A_3]g_{H2O}(r)$	9	89	0.22
(7e) RTN decomposition, $[A_1, A_2, A_3]g_{H2O}(r), \Delta T$	10	86	0.21

For reference, we list the values for a gravity-only model of the trajectory in addition to those for the different non-gravitational models. In addition to a model involving an impulsive change in velocity, we consider continuous non-gravitational accelerations $g(r)$ with a dependence on the heliocentric distance r that is either a power law or, for H₂O or CO volatiles (g_{H2O} or g_{CO}), based on cometary outgassing models^{8,30}. The acceleration vector can be inertially fixed or decomposed in either the radial, transverse, normal (RTN; components indicated as $A_1 g(r)$, $A_2 g(r)$ and $A_3 g(r)$, respectively) or the along-track, cross-track, normal (ACN; components indicated as $A_{AG}(r)$, $A_{AC}(r)$ and $A_{AN}(r)$, respectively) frame. We also test the possibility of a time delay ΔT with respect to perihelion for the peak of the outgassing activity. The numbering of the models refers to the discussion in Methods.

in an HST-only arc complemented with just a few early ground-based high-quality data points.

Exploring various possible explanations for the non-gravitational acceleration that was detected, we find outgassing to be the most physically plausible explanation, although with some caveats. A thermal outgassing model¹⁵, which treats 'Oumuamua like a common cometary nucleus, suggests a non-gravitational force proportional to r^{-2} in the range of distances covered by our observations.

The model predictions for the magnitude and temporal evolution of the non-gravitational acceleration are within a factor of about 2–3 of the observations (see Methods) for a water production rate of $Q_{H2O} = 4.9 \times 10^{25}$ molecules s⁻¹ (or 1.5 kg s^{-1}) near 1.4 AU and an additional contribution from $Q_{CO} = 4.5 \times 10^{25}$ molecules s⁻¹ (or 2.1 kg s^{-1}). Outgassing at this level does not conflict with the absence of spectroscopic detections for outgassing of OH, because the values quoted are well below the spectroscopic limits on production rates¹⁶. However, the inferred upper limits for water production at 1.4 AU, which are based on the non-detection of CN⁷ and assumed Solar System abundances for Q_{CN}/Q_{OH} ¹⁷, show that 'Oumuamua would need to be substantially depleted in CN (by a factor of more than about 15) relative to water. The model also predicts 0.4 kg s^{-1} of dust production, which should have been detectable in the images. However, if the grains are predominantly larger than a few hundred micrometres to millimetres, they would not have been detected at optical wavelengths (see Methods). In the Solar System, comet 2P/Encke is noteworthy for its lack of small dust near perihelion¹⁸. Cometary behaviour implies that 'Oumuamua must have some internal strength, at least comparable to Solar System comets¹⁹, because asteroid-like densities are ruled out (see Methods).

Alternative explanations for the observed acceleration include (1) solar-radiation pressure, (2) the Yarkovsky effect, (3) friction-like effects aligned with the velocity vector, (4) an impulsive change in velocity, (5) a binary or fragmented object, (6) a photocentre offset or (7) a magnetized object. However, as outlined in the following, these explanations are all either physically unrealistic or insufficient to explain the observed behaviour.

(1) The simplest physical phenomenon that could cause a radial acceleration that follows an r^{-2} dependence and that is directed away

from the Sun is pressure from solar radiation. Such a pressure has been detected for a few small asteroids^{20–23}; however, for 'Oumuamua the magnitude of the observed acceleration implies an unreasonably low bulk density, roughly three to four orders of magnitude below the typical density of Solar System asteroids of comparable size. Additional considerations regarding the plausibility of solar-radiation pressure as an explanation for the non-gravitational motion are presented in Methods.

(2) A rotating body in space experiences a small force due to the anisotropic emission of thermal photons, the so-called Yarkovsky effect²⁴. The resulting perturbation can be excluded as an explanation for the observed acceleration because of its low intensity (at most comparable to that of solar-radiation pressure) and because it mainly affects the motion in the along-track direction, in conflict with our data.

(3) Some dynamical effects, such as friction- or drag-like phenomena, tend to be aligned with the direction of motion and not with the heliocentric radial vector. However, decomposition of the non-gravitational acceleration shows that the respective best-fitting component along the direction of motion is not only insufficient to explain the observations (see Table 1) but also positive, whereas drag-like phenomena would require it to be negative.

(4) Models of the trajectory that include a single impulsive change in velocity, for example, due to a collision, provide a poorer fit to the data (Table 1) than purely radial acceleration. More importantly, we detect the non-gravitational signal even in disjoint subsets of the observed arc, separated at the time of the possible impulse, which makes continuous acceleration a far more likely explanation.

(5) In the case of a binary or fragmented object, the centre of mass of the combined system does in fact follow a purely gravitational trajectory, and the detected non-gravitational signature could be an artefact caused by us tracking only the main component of 'Oumuamua. However, no secondary body or fragment is visible in our data down to a few magnitudes fainter than 'Oumuamua, and any object smaller than the corresponding size limit (roughly 100 times smaller than 'Oumuamua) would be insufficient to explain the observed astrometric offsets.

(6) 'Oumuamua may have surface characteristics that significantly displace the optical photocentre (the position that is measured astrometrically) from the centre of mass. However, even assuming the longest possible extent of 800 m for a low-albedo ($p = 0.04$) object¹, the maximum separation between the two reference points would be approximately $0.005''$ at closest approach, several orders of magnitude less than the offset observed for a gravity-only solution.

(7) If 'Oumuamua had a strong magnetic field, then interaction with solar wind could affect its motion^{25,26}. Assuming a dipole field, a plasma-fluid model and typical solar wind speed and proton number density²⁷, we find the resulting acceleration for an object of the nominal size of 'Oumuamua¹ to be only $2 \times 10^{-11} \text{ m s}^{-2}$, too small by a factor of about 10^5 , even if we adopt the high magnetization and density of asteroid (9969) Braille²⁸.

Although this list of possible alternative explanations is not exhaustive, we believe that it covers most of the physical mechanisms worth exploring on the basis of the available data. The models tested here attempt only to describe the dynamical behaviour of 'Oumuamua within the temporal arc covered by the observations. The presence of non-gravitational acceleration and the complexity of the physical explanation proposed by us suggest that any extrapolation to the past and future trajectories of 'Oumuamua outside the modelled arc may be subject to substantial uncertainties.

Outgassing provides the most plausible physical model of the non-gravitational acceleration by postulating that 'Oumuamua behaves like a miniature comet. This hypothesis is consistent with independent results^{5,6} that demonstrate that the spectra and the lack of activity observed are consistent with a cometary body with a thin insulating mantle, and also with the non-gravitational accelerations observed for other Solar System comets (see Extended Data Fig. 1). By establishing the object as an icy body (albeit one with possibly unusual chemical

composition and dust properties), this scenario agrees with the predictions that suggest that only a small fraction of interstellar objects should be asteroidal²⁹. The lack of observed dust lifted from the object by the hypothesized cometary activity can be explained by an atypical dust-grain size distribution that is devoid of small grains, a low dust-to-ice ratio or surface evolution from its long journey. However, these important aspects of the physical nature of 'Oumuamua cannot be resolved conclusively with the existing observations. In situ observations would be essential to reveal unambiguously the nature, origin and physical properties of 'Oumuamua and other interstellar objects that may be discovered in the future. This work shows that although 'Oumuamua looks familiar there are differences that relate to its birth in a solar system far from our own.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0254-4>.

Received: 17 April 2018; Accepted: 29 May 2018;

Published online 27 June 2018.

1. Meech, K. J. et al. A brief visit from a red and extremely elongated interstellar asteroid. *Nature* **552**, 378–381 (2017).
2. Fraser, W. C. et al. The tumbling rotational state of 1I/'Oumuamua. *Nature Astron.* **2**, 383–386 (2018).
3. Drahus, M. et al. Tumbling motion of 1I/'Oumuamua reveals body's violent past. *Nature Astron.* **2**, 407–412 (2018).
4. Belton, M. J. S. et al. The excited spin state of 1I/2017 U1 'Oumuamua. *Astrophys. J.* **856**, L21 (2018).
5. Fitzsimmons, A. et al. Spectroscopy and thermal modelling of the first interstellar object 1I/2017 U1 'Oumuamua. *Nature Astron.* **2**, 133–137 (2018).
6. Jewitt, D. et al. Interstellar Interloper 1I/2017 U1: observations from the NOT and WIYN Telescopes. *Astrophys. J.* **850**, L36 (2017).
7. Ye, Q.-Z., Zhang, Q., Kelley, M. S. P. & Brown, P. G. 1I/2017 U1 ('Oumuamua) is hot: imaging, spectroscopy, and search of meteor activity. *Astrophys. J.* **851**, L5 (2017).
8. Marsden, B. G., Sekanina, Z. & Yeomans, D. K. Comets and nongravitational forces. *V. Astron. J.* **78**, 211–225 (1973).
9. Królikowska, M. Long-period comets with non-gravitational effects. *Astron. Astrophys.* **427**, 1117–1126 (2004).
10. Wainscoat, R. et al. The Pan-STARRS search for near earth objects. *Proc. IAU* **10**, 293–298 (2015).
11. Denneau, L. et al. The Pan-STARRS moving object processing system. *Publ. Astron. Soc. Pacif.* **125**, 357–395 (2013).
12. Williams, G.V. MPEC 2017-U181: comet C/2017 U1 (PANSTARRS). *IAU Minor Planet Center* <https://minorplanetcenter.net/mpec/K17/K17U11.html> (2017).
13. Engelhardt, T. et al. An observational upper limit on the interstellar number density of asteroids and comets. *Astron. J.* **153**, 133 (2017).
14. Farnocchia, D., Chesley, S. R., Milani, A., Gronchi, G. F. & Chodas, P. W. in *Asteroids IV* (eds Michel, P. et al.) 815–834 (Univ. Arizona Press, Tuscan, 2015).
15. Prialnik, D. Modeling the comet nucleus interior. *Earth Moon Planets* **89**, 27–52 (2000).
16. Park, R. S., Pisano, D. J., Lazio, T. J. W., Chodas, P. W. & Naidu, S. P. Search for OH 18-cm radio emission from 1I/2017 U1 with the Green Bank Telescope. *Astron. J.* **155**, 185 (2018).
17. Cochran, A. L., Barker, E. S. & Gray, C. L. Thirty years of cometary spectroscopy from McDonald Observatory. *Icarus* **218**, 144–168 (2012).
18. Fink, U. A taxonomic survey of comet composition 1985–2004 using CCD spectroscopy. *Icarus* **201**, 311–334 (2009).
19. McNeill, A., Trilling, D. E. & Mommert, M. Constraints on the density and internal strength of 1I/'Oumuamua. *Astrophys. J.* **857**, L1 (2018).
20. Williams, G.V. MPEC 2008-D12: 2006 RH120. *IAU Minor Planet Center* <https://minorplanetcenter.net/mpec/K08/K08D12.html> (2008).
21. Micheli, M., Tholen, D. J. & Elliott, G. T. Detection of radiation pressure acting on 2009 BD. *New Astron.* **17**, 446–452 (2012).
22. Micheli, M., Tholen, D. J. & Elliott, G. T. 2012 LA, an optimal astrometric target for radiation pressure detection. *Icarus* **226**, 251–255 (2013).
23. Micheli, M., Tholen, D. J. & Elliott, G. T. Radiation pressure detection and density estimate for 2011 MD. *Astrophys. J.* **788**, L1 (2014).
24. Vokrouhlický, D., Bottke, W. F., Chesley, S. R., Scheeres, D. J. & Statler, T. S. in *Asteroids IV* (eds Michel, P. et al.) 509–531 (Univ. Arizona Press, Tuscan, 2015).

25. Meyer-Vernet, N. *Basics of the Solar Wind* 348–351, 366–371 (Cambridge Univ. Press, Cambridge, 2007).
26. Zubrin, R. M. & Andrews, D. G. Magnetic Sails and Interplanetary Travel. *J. Spacecr. Rockets* **28**, 197–203 (1991).
27. Wang-Sheeley-Arge (WSA)-Enlil Solar Wind Prediction. *Space Weather Prediction Center* <https://www.ngdc.noaa.gov/enlil/> (accessed March 2018).
28. Richter, I. et al. Magnetic field measurements during the ROSETTA flyby at asteroid (21) Lutetia. *Planet. Space Sci.* **66**, 155–164 (2012).
29. Meech, K. J. et al. Inner solar system material discovered in the Oort cloud. *Sci. Adv.* **2**, e1600038 (2016).
30. Yabushita, S. On the effect of non-gravitational processes on the dynamics of nearly parabolic comets. *Mon. Not. R. Astron. Soc.* **283**, 347–352 (1996).

Acknowledgements K.J.M., J.T.K. and J.V.K. acknowledge support through NSF awards AST1413736 and AST1617015, in addition to support for HST programmes GO/DD-15405 and -15447 provided by NASA through a grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy under NASA contract NAS 5-26555. R.J.W. and R.W. acknowledge support through NASA under grant NNX14AM74G issued to support Pan-STARRS1 through the SSO Near Earth Object Observation Program. D.F., P.W.C. and A.E.P. conducted this research at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. We thank S. Sheppard for obtaining the Magellan observations, and E. J. Christensen, W. H. Ryan and M. Mommert for providing astrometric uncertainty information related to the Catalina Sky Survey, Magdalena Ridge Observatory and Discovery Channel Telescope observations of 'Oumuamua. This work is based on observations obtained at CFHT, which is operated by the National Research Council of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique of France and the University of Hawai'i. It is based in part on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programme 2100.C-5008(A) and in part on observations obtained under programme GS-2017B-DD-7 obtained at the Gemini Observatory, which is operated by AURA under cooperative agreement with the NSF on behalf of the Gemini partnership: NSF (United States), NRC (Canada), CONICYT (Chile), MINCYT (Argentina) and MCT (Brazil). This work is also based on observations made with NASA/ESA HST, obtained at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy under NASA contract NAS 5-26555. This work has made use of data from the ESA mission Gaia (<https://www.cosmos.esa.int/gaia>), processed by the Gaia Data Processing and Analysis Consortium (DPAC; <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

Reviewer information *Nature* thanks A. Fitzsimmons, M. Granvik and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.M. discovered the non-gravitational acceleration and extracted the high-precision astrometry from most ground-based observations obtained by the team. D.F. performed the different fits and modelling of the non-gravitational acceleration. K.J.M. secured the HST time and designed the observation programme, computed sublimation dust and gas outgassing limits, and provided the assessment of outgassing. M.W.B. led the design of the HST observations and contributed precision astrometry from HST images. O.R.H. obtained the deep stack of images, searched them for dust and companion, and estimated production rates. D.P. performed the thermal sublimation modelling. N.S. conducted thermal model calculations. H.A.W. managed the HST observations and the initial reduction of images. P.W.C. provided support in analysing possible explanations for the observed non-gravitational acceleration. J.T.K. assembled the deep stack of CFHT data to search for dust and outgassing. R.W. identified and searched pre-discovery images of 'Oumuamua in Pan-STARRS1 data. R.J.W. obtained the observations using CFHT and searched for pre-discovery observations of 'Oumuamua. H.E. contributed to the HST proposal and to the design of the HST observations. J.V.K. and K.C.C. contributed to the HST proposal. D.K. provided support in analysing possible explanations for the observed non-gravitational acceleration. A.E.P. investigated the magnetic hypothesis.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0254-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Ground-based observations. We found the first evidence of non-gravitational forces acting on 'Oumuamua in astrometry derived from a set of ground-based optical images obtained with various ground-based telescopes¹. Our first optical follow-up observations were performed with ESA's 1.0-metre Optical Ground Station (OGS) in Tenerife, Spain, only 13 h after the discovery of 'Oumuamua. Subsequent deeper observations were conducted with the 3.6-metre CFHT (seven nights), the 8.2-metre ESO VLT Unit Telescope 1 (two nights), and the 6.5-metre Magellan Baade telescope (two nights). The astrometric positions derived from this ground-based dataset, together with the associated error bars, are sufficient to detect the non-gravitational acceleration at a level of significance of about 5σ .

Search for pre-discovery detections. We searched for pre-discovery images of 'Oumuamua at positions computed from a model trajectory that included the observed non-gravitational acceleration. Pan-STARRS1 observed suitable fields through its broad *w*-band filter on 2017 June 18 and 22 and through its *i*-band filter on 2017 June 17, almost three months before perihelion. During this time, the predicted average brightness of 'Oumuamua was around $V \approx 26$ (uncertain because of the large amplitude of the light curve of the object), much fainter than the limiting magnitude of Pan-STARRS1. No object was visible in these images at the predicted location.

HST data and astrometry. Images of 'Oumuamua were obtained with HST in two separate awards of Director's Discretionary (DD) time. The first set of observations was designed soon after the discovery of 'Oumuamua, with the primary goal of extending the observational arc to obtain tighter astrometric constraints on the trajectory of the object. Three HST visits were executed on 2017 November 21–22, one visit on 2017 December 12 and a fifth on 2018 January 2. To maximize the length of the orbital arc covered, the last observation was set to be performed as late as possible, assuming that we would know the rotational phase sufficiently well to enable us to catch the steadily fading and only barely detectable target at light-curve maximum. The discovery of non-principal-axis rotation^{2–4} invalidated our assumption of a predictable light curve and motivated a second allocation of four additional HST orbits, added to the final visit, that allowed us to cover 'Oumuamua in a more sophisticated temporal cadence that was designed to maximize its detectability regardless of light-curve phase. This additional allocation was essential for our final detection.

Each visit used the same basic observing pattern of five 370-second exposures of the full field of WFC3/UVIS, an exposure time that is just long enough to accommodate CCD readout and data-storage overheads without loss of integration time within the allocated single orbit. All images were taken through the extremely broad F350LP filter, chosen for maximum throughput. This strategy was modelled after very similar observations of (486958) 2014 MU₆₉, the target of the New Horizons extended mission, and resulted in a signal-to-noise ratio of approximately 2–3 for a solar-colour object of magnitude $R = 27.5$.

During all observations, HST tracked 'Oumuamua, and target motions and parallax corrections were applied. As a result, the object appears as a point source in our images and the background-field stars appear as long trails. Because the density of background stars was very low for these observations, the exact placement of our target within the field of view of the instrument had to be adjusted for some visits to ensure that the number of reference stars (3–10) was sufficient for the aimed-at high-precision astrometric solution.

The positions of reference stars were determined from PSF fitting using the Tiny Tim model³¹ and application of a smearing function derived from the HST-centric motion of the object during each exposure. Uncertainties of the resulting position and flux measurements were derived using a Markov chain Monte Carlo sampling algorithm³². The probability density functions (PDFs) from this calculation were then used to update the default world coordinate system (WCS) solution of each image, using the Gaia DR2³³ position of each star as a reference. A PDF was also derived for this final reference WCS.

The position of 'Oumuamua was computed in the same fashion, except that no smearing function was needed. Object position, flux and a PDF were derived for each frame where possible (a few images were lost to cosmic-ray strikes). In the final visit, our target was detected in only two of the five orbits. Using the WCS PDF for reference, we combined these results to obtain the final sky-plane PDF for the object in each image and then converted the PDF to a Gaussian approximation covariance for use in the fitting of the trajectory of 'Oumuamua. Whereas the resulting uncertainties are dominated by catalogue errors for the earlier visits, the low signal-to-noise ratio of the object contributes substantially to the error budget for the final visit. The formal uncertainties from this procedure are at most $0.01''$ – $0.02''$.

Accumulated observational dataset. Our attempts to constrain the trajectory of 'Oumuamua made use of all available astrometric positions. In addition to our own astrometric dataset (see Extended Data Tables 1 and 2), we included all relevant data submitted to the Minor Planet Center, for a total of 177 ground-based observations and 30 HST observations. Seven additional ground-based observations

deemed unreliable by the observers were not considered. Where no uncertainties were provided by the observers, we assumed a $1''$ positional uncertainty, except for a handful of observations that showed poor internal consistency were further de-weighted (these error bars are presented in Extended Data Table 3). Moreover, we assumed that the reported observation times are uncertain by 1 s. Finally, positions that did not use the Gaia DR1 or DR2 catalogue^{33,34} as a reference were corrected for systematic errors of the respective star catalogue³⁵, resulting in corrections as large as $0.4''$ for the USNO-B1.0 catalogue³⁶. To mitigate the effect of unresolved systematic errors, we used an uncertainty floor of $0.05''$ to set the data weights.

Potential biases in the detection of non-gravitational motion. To test whether the detected non-gravitational acceleration could be an artefact introduced by a subset of biased astrometric observations, we used the $A_{1g}(r), g(r) \propto r^{-2}$ non-gravitational model. We performed a series of analyses on subsets of the full data arc that were designed to highlight whether specific groups of observations could be responsible for the signal. A summary of our findings is as follows.

The signal is not caused by the early, noisier observations. Fitting only data taken after 2017 October 25 or after 2017 November 15 still yields a detection of A_1 at 17σ and 2.5σ confidence, respectively. Similarly, the signal is not caused by only the late part of the arc. Fitting only data taken before 2017 November 15 or up to 2017 December 1 still yields a detection of A_1 at 2.8σ and 7.4σ confidence, respectively.

To rule out biases in data from ground-based observations, for example, due to colour refraction in the atmosphere, we computed orbital solutions using only HST data and a single ground-based observation set: OGS on October 19, CFHT on October 22 or VLT on October 25. In all three tests, non-gravitational motion was detected at a significance of at least 11σ .

Some of the ground-based astrometric positions for 'Oumuamua were measured relative to the Gaia DR1 catalogue, which does not include the proper motions of stars. Because Gaia DR1 uses 2015 as the reference epoch, offsets due to proper motions³⁵ could amount to as much as about $0.04''$. The tests that we performed that combine HST and our ground-based astrometry, which was reduced with Gaia DR2, shows that the detection of non-gravitational motion is not caused by this issue.

To rule out the possibility that the detection of non-gravitational motion could be due to issues with HST data (such as in the case of comet C/2013 A1, for which the HST astrometry was found to have larger errors than expected³⁷), we performed a fit using only ground-based observations and still detected non-gravitational motion at 7.1σ significance. To make sure that the high significance of the non-gravitational signal is not caused by overly optimistic assumptions regarding the astrometric uncertainties, we ran a test using an uncertainty floor of $1''$ and still obtained a 7.0σ signal for A_1 .

The results of our tests show that the observed non-gravitational signature is not an artefact of biases in the data or the specifics of the analysis performed, but is indeed present in the motion of 'Oumuamua.

Non-gravitational models. In addition to $A_{1g}(r)$, with $g(r) \propto r^{-2}$, we considered several alternative models for the observed non-gravitational acceleration of 'Oumuamua. The χ^2 and χ^2_ν values of the corresponding fits to all astrometric data are shown in Table 1 for comparison with the gravity-only reference model. A brief summary of each model (numbered as in Table 1) is provided below.

(1) We searched for evidence of an impulsive change in velocity (Δv) and found two χ^2 minima, one on 2017 November 5 and another on 2017 December 6, both requiring $\Delta v \geq 5 \text{ m s}^{-1}$. However, the corresponding orbital solutions provide a poorer fit to the data than do continuous acceleration models. Moreover, as discussed before, evidence of non-gravitational acceleration is found in the arcs before 2017 December 6 and after 2017 November 5. Therefore, an impulsive Δv event alone cannot model the trajectory of 'Oumuamua.

(2) We tested different power laws for the radial dependency of the acceleration: $g(r) \propto r^{-k}$, $k \in \{0, 1, 2, 3\}$. A constant $g(r)$ ($k=0$) provides a poorer fit to the data. Within the timespan of our fit, which extends from $r = 1.1 \text{ AU}$ to $r = 2.9 \text{ AU}$, the acceleration decreases with increasing heliocentric distances at a rate that cannot be much steeper than r^{-2} , but can be gentler, for example, r^{-1} , with both trends having comparable likelihood. On the other hand, a trend of r^{-3} is strongly disfavoured by the data.

(3) Adding transverse ($A_{2g}(r)$) and normal (out-of-plane; $A_{3g}(r)$) acceleration components to a radial-acceleration-only model (the result is referred to as the RTN model) yields only a modest improvement in the fit, regardless of the dependence on heliocentric distance that we select, showing that the non-gravitational acceleration of 'Oumuamua is mostly radial. The best-fitting values for A_2 and A_3 are consistent with zero (significance of less than 1σ) and are an order of magnitude smaller than that for A_1 .

(4) Alternatively, the acceleration vector can be decomposed into along-track, cross-track, and normal (ACN) components with respect to the trajectory. The goodness of the resulting fit is comparable to that obtained for the RTN decomposition. However, in the ACN frame all three directions are needed to describe the data, whereas a single parameter is sufficient in the RTN frame.

(5) An unacceptably poor fit is obtained if the acceleration is assumed to act exclusively in the direction of the velocity vector of the object (that is, the along-track component of the ACN frame), for any $g(r) \propto r^{-k}$, $k \in \{0, 1, 2, 3\}$.

(6) We also tested the possibility of a constant acceleration vector, fixed in inertial space. Despite the larger number of estimated parameters, the resulting fit is no better than that obtained with a purely radial acceleration. Moreover, the complex rotation state of 'Oumuamua^{2–4} is at odds with such an inertially fixed acceleration.

(7) Finally, we tested non-gravitational models involving cometary activity. A CO-driven³⁰ $g(r)$ (7a and 7c in Table 1) behaves similarly to r^{-2} for $r < 5$ AU and provides a better fit than a H₂O-driven⁸ $g(r)$ (7b and 7d), which falls off like $r^{-2.15}$ for $r < 2.8$ AU and then decays abruptly like r^{-26} . This latter model can include a time offset $\Delta T = 55$ d with respect to perihelion for the acceleration peak³⁸ (7e), thus moving the fast decay of $g(r)$ outside of the data arc.

The difference between χ^2 values for models within a given family (the exponent k for each of models (2), (3), (4) and (5) in Table 1) is useful for statistically evaluating how significantly some exponents are disfavoured with respect to the best-fitting one of the same family.

Limits on cometary activity. We estimate that no more than about 1 kg of 1-μm-sized dust grains could have been present in the direct vicinity of 'Oumuamua (less than 2.5'' or 750 km from the nucleus) on 2017 October 25–26¹, on the basis of the dust-limiting magnitude for dust $g > 29.8$ mag arcsec⁻². Here we perform the same analysis on deep stacks of the 2017 November 21, 22 and December 1 HST data in search of evidence of dust. To this end, we subtracted a copy of each image from itself after rotation by 180°. Because any dust is pushed from the nucleus by solar-radiation pressure, its distribution is expected to be highly asymmetric. The self-subtraction removes the light from the nucleus and the symmetric component and makes the asymmetric component more prominent. The subtracted frames were further enhanced by wavelet filtering (which boosts the signal with spatial frequencies corresponding to 2–8 pixels) and adaptive smoothing (which smooths the signal over a region with a size adapted dynamically such that the signal-to-noise ratio reaches a threshold, set here to 2). Examination of the resulting images, shown in Fig. 1, does not reveal any sign of dust to a similar limit. The asymmetry test is particularly sensitive for the October 25–26 stack: because the Earth was only 15° above the orbital plane of the object, any dust released from the nucleus since its passage through perihelion is expected to be confined to a narrowly fanning region with position angles of approximately 96°–135°. Our findings thus indicate that the original upper limit of about 1 kg of 1-μm-sized dust within 750 km on October 25 is conservative (corresponding to $g > 29.8$ mag arcsec⁻² at the 5σ level).

To test this limit, a dust feature was introduced in the images, which were then re-processed using the same enhancement techniques. The feature was produced using a cometary image approximately matching the expected morphology of ejected dust for October 25 (when the geometry was the best to concentrate the dust in a narrow region), scaled to match the photometric contribution in the central 2.5''. This is illustrated in Fig. 1, and indicates that the dust would indeed probably be detected.

From the orbital fits we know that the non-gravitational acceleration on 'Oumuamua on October 25 at $r = 1.4$ AU was $A_1 r^{-2} = 2.7 \times 10^{-6}$ m s⁻¹. The mass m of 'Oumuamua can be estimated from the photometry¹, assuming an albedo of 0.04 (0.2) and a bulk density of less than 500 kg m⁻³ (2000 kg m⁻³) for a cometary³⁹ (asteroidal^{19,40}) object. If the non-gravitational force is due to cometary activity, then Newton's law can be used to relate the observed acceleration to the gas production rate⁴¹ Q : $ma = Q\zeta v_i$, where v_i is the gas ejection velocity and ζ is a poorly constrained, dimensionless efficiency factor that accounts for (among other effects) the geometry of the emission. At the heliocentric distance of 'Oumuamua on October 25 of 1.4 AU, ζv_i would fall between 150 m s⁻¹ and 450 m s⁻¹; in the following, we adopt 300 m s⁻¹. The resulting gas production rates, at a heliocentric distance of 1.4 AU, range from 0.7 kg s⁻¹ to 140 kg s⁻¹ depending of the size, shape and mass of the object, with a mass loss of $Q = 10$ kg s⁻¹ being our best estimate. This value was used to constrain the thermal model discussed in the following.

Thermal model. We carried out thermal model calculations to estimate the interior temperatures that 'Oumuamua reached during its passage. These thermal calculations begin four years before perihelion and end two years after perihelion. The one-dimensional⁴² model resolves the diurnal cycle with at least 288 time steps within each 7.34-hour simple rotation. We assumed an albedo of 0.04 and an obliquity of 45°, and used two parameter combinations: one with a porosity of 40% and a thermal inertia of 400 J m⁻² K⁻¹ s^{-1/2} (at 200 K), and the other with a porosity of 90% and a thermal inertia of 40 J m⁻² K⁻¹ s^{-1/2}. Calculations were carried out for the object's equator (where the surface normal is perpendicular to the rotation axis) and at a latitude of 45°, starting from an initial temperature of 4 K. The depths to maximum temperature along the orbit depend on the assumed physical properties, but for the parameters specified above, which capture a wide range of values, 160 K (the approximate activation threshold for H₂O-driven cometary activity) is reached within the top roughly 1 m of the surface, consistent

with previous results⁵. Because 'Oumuamua is only tens of metres wide, 30 K (the approximate threshold for CO activity) was exceeded within most of the body. The case of CO₂ lies in between (80 K). The model temperatures suggest that if CO ice was present then considerable outgassing occurred, and even CO₂ ice would have experienced substantial sublimation.

Outgassing models. To verify whether cometary activity can produce the observed non-gravitational acceleration, we modelled¹⁵ the object as a comet. Note that, because of the large range of plausible masses for the nucleus, our results should be considered order-of-magnitude estimates. We assumed the following initial physical characteristics for a spherical nucleus¹: a radius of 102 m, an albedo $\rho = 0.04$, a density $\rho = 500$ kg m⁻³, an ice-to-dust ratio of unity (in mass), 60% porosity and a bulk thermal conductivity of 0.7 W m⁻¹ K⁻¹, all typical values for comets¹⁵. The model considers subsurface H₂O and CO ices (with CO/H₂O = 0.05 by mass) and, following this model nucleus along the trajectory of 'Oumuamua, evaluates the sublimation over a 400-day period centred on perihelion. The water production rate was found to peak close to perihelion and then decline following an r^{-2} profile until 100 days after perihelion (at 2.6 AU in mid-December 2017), when it starts to decrease sharply. At that point, the CO production rate, which does not change much along the trajectory, becomes dominant, and hence the total production rate continues to follow the r^{-2} trend. The gas velocity was estimated at $v_i = 500$ m s⁻¹, within the range of ζv_i values discussed above.

We adjusted additional physical parameters that characterize the model nucleus (such as thermal conductivity, ice-to-dust ratio and bulk density) in an attempt to match $Q_{H_2O} = 10$ kg s⁻¹ at 1.4 AU, our estimate of the gas production rate required to generate the observed non-gravitational acceleration. The closest match to the observations resulted from the following model parameters: $\rho = 450$ kg m⁻³, ice/dust = 3 by mass, CO/H₂O = 0.25, 60% porosity for the initial composition and low temperature. The resulting model parameters are mostly within acceptable limits and physically meaningful; for instance, the thermal conductivity required matches that of silicates, rather than that of a mix of silicate and organics. The dust production was estimated using a low drag coefficient, acknowledging that the gas, and therefore the dust, would come from the subsurface. For our initial model, however, $Q_{dust} = 0.2$ kg s⁻¹ and the maximal gas production at 1.4 AU is $Q_{H_2O} = 2.5$ kg s⁻¹, which provides insufficient acceleration. With a much higher CO/H₂O ice ratio, the production rate increases to within a factor of about 2–3 that needed to match the acceleration detected, with a dust production rate of 0.4 kg s⁻¹. A further increase in mass loss by approximately 30% would result if the surface area had an ellipsoidal shape instead of a spherical shape, with the same median photometric cross-section. The dust production rates inferred from the thermal models require the grains to be relatively large (about 100 μm to a few millimetres) to match the optical non-detection limits for dust. Large grains are typical of outgassing from subsurface layers as seen in laboratory experiments⁴³, and models of the physical interaction of Oort cloud comets and the interstellar medium show that small grains are efficiently removed by drag effects⁴⁴. No model using an asteroid-like density¹⁹ could be made to produce sufficient acceleration. Further, a high bulk density imposes a limit on ice content even for near-zero porosity. Even assuming a very high CO/H₂O ratio, the maximum outgassing is more than an order of magnitude too low. Finally, acceleration from outgassing reaches the required value if the assumed density of 'Oumuamua is lowered to around 200 kg m⁻³. Although other values could be obtained by adjusting the dust size distribution and the nucleus pore size, further exercises would be of little benefit, as long as we do not have additional constraints.

In conclusion, we find that sublimation can account for the non-gravitational forces that were measured, when modelling 'Oumuamua as a small comet, but only if it has some unusual properties.

Consequences of the analysis for the study of the origin of 'Oumuamua. The many uncertainties and assumptions in the non-gravitational models presented here limit our ability to fully determine the past history of 'Oumuamua. These limitations are intrinsically due to the absence of observational information on the behaviour of the non-gravitational acceleration outside the observed arc. In particular, the absence of information on the behaviour of the non-gravitational acceleration before the time of discovery implies that it is much more difficult (and subject to much larger uncertainties) to extrapolate the motion of 'Oumuamua to its original incoming direction.

Solar-radiation pressure. A simple radial dependency of the non-gravitational acceleration, decaying as r^{-2} with the heliocentric distance, is allowed by the dataset for $A_1 = (4.92 \pm 0.16) \times 10^{-6}$ m s⁻². If interpreted as solar-radiation pressure on the projected area of the object exposed to sunlight, then this A_1 value would correspond to an area-to-mass ratio between about 0.5 m² kg⁻¹ and 1 m² kg⁻¹. Given the range of possible sizes and shapes of 'Oumuamua¹, and assuming a uniform density and an ellipsoidal shape for the body, this estimate of the area-to-mass ratio would correspond to a bulk density of the object between about 0.1 kg m⁻³ and 1 kg m⁻³, three to four orders of magnitude less than that of water. Alternatively, to be composed of materials with densities comparable to normal asteroidal or cometary

matter, 'Oumuamua would need to be a layer, or a shell, at most a few millimetres thick, which is not physically plausible.

Unless 'Oumuamua has physical properties that differ markedly from those of typical Solar System bodies within the same size range, the interpretation of the non-gravitational acceleration being due to solar-radiation pressure is therefore unlikely.

Binary object or fragmentation event. The existence of one or more fragments could theoretically explain the detected astrometric offsets by displacing the centre of mass of the overall system from the main component that was measured astrometrically. However, the existence of a bound secondary body of substantial mass can easily be discounted both directly and indirectly.

The offsets from a gravity-only solution (see Fig. 2) observed at the time of our deepest images are at the arcsecond level, requiring a possible bound secondary body to have a separation from the main mass that is comparable or greater than this distance. However, no co-moving object was detected in the vicinity of the main body, although most of the images we obtained with large-aperture telescopes have subarcsecond resolution and reach a depth a few magnitudes fainter than 'Oumuamua. Specifically, the limiting magnitudes estimated from the signal-to-noise ratio of 'Oumuamua on deep stacks of data from VLT (October 25), GS (October 26) and HST (November 21 and 22) are $r'_{\text{lim}} = 27.0$ and $V_{\text{lim}} = 29.2$, respectively. Conversion to an upper limit for the radius of an unseen object yields 7.8 m (3.5 m) and 4.5 m (2.0 m), respectively, for an albedo of 0.04 (0.2), typical for a cometary (an asteroid) nucleus—about 100 times smaller than the main body using the same assumptions. In addition, given the small mass of 'Oumuamua, the radius of its sphere of influence $r \propto a(m/M)^{2/5}$ (where a is the distance between the object and the Sun and m and M the masses of the object and the Sun, respectively) is of the order of 1 km, corresponding to angular separations of milliarcseconds. Any object within such a distance would be fully embedded in the PSF of the main body and therefore would not contribute any detectable offset to the astrometric photocentre.

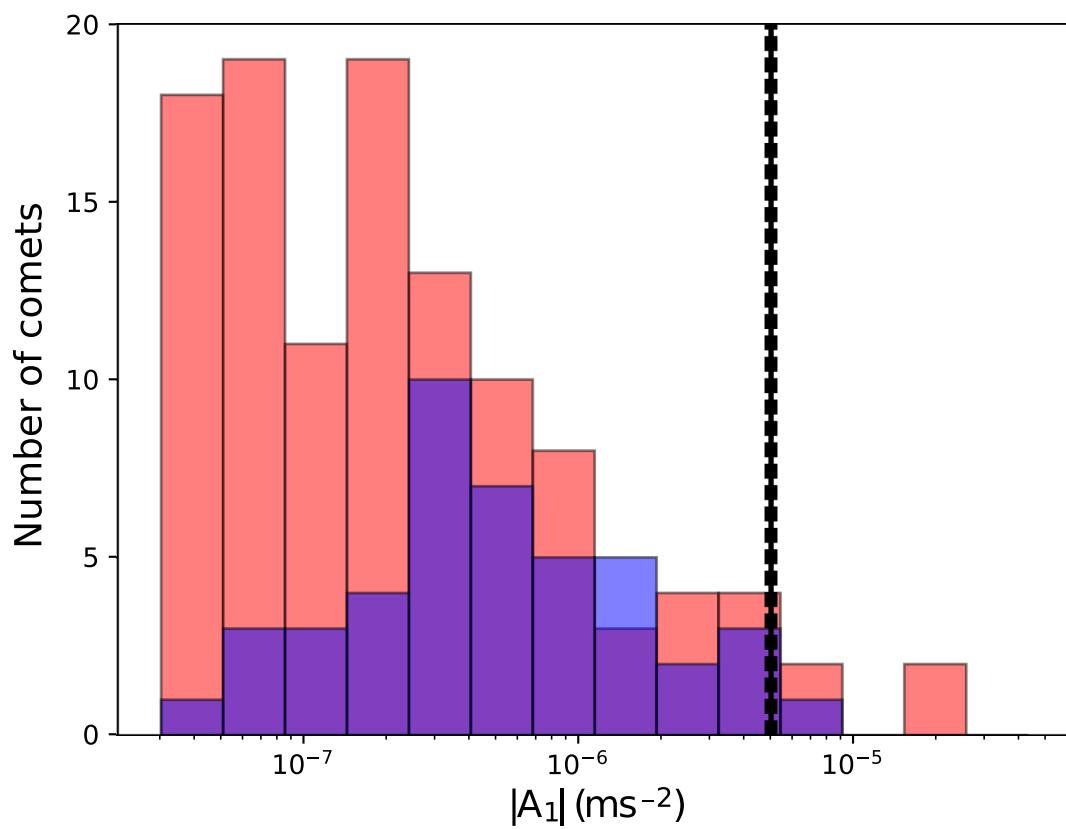
The possibility of an unbound fragment being ejected by 'Oumuamua during the observed arc can also be excluded, not only because no such fragment was seen in the deep images we obtained, but also because its dynamical effect would correspond to an impulse-like event in the trajectory, which we have already shown to be incompatible with the data.

Code availability. The JPL asteroid and comet orbit determination code, used in the in-depth analysis of the possible dynamical scenarios, is proprietary. However, some key results of this analysis, including the detection of a significant non-gravitational acceleration at the 30σ level, can easily be reproduced by using freely

available software such as Find_Orb (https://www.projectpluto.com/find_orb.htm). The code for the comet sublimation model is a direct implementation of a published model^{15,45}. Source code and further documentation for the type of one-dimensional thermal model used is available at <https://github.com/nschorgh/Planetary-Code-Collection/>.

Data availability. The astrometric positions and uncertainties on which this analysis is based are available in Extended Data Tables 1–3, and will be submitted to the Minor Planet Center for public distribution. Source Data for Fig. 2 and Extended Data Fig. 1 is available with the online version of the paper.

31. Krist, J. E., Hook, R. N. & Stoehr, F. 20 years of Hubble Space Telescope optical modeling using Tiny Tim. *Proc. SPIE* **8127**, 81270J (2011).
32. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
33. Gaia Collaboration. Gaia data release 2. Summary of the contents and survey properties. *Astron. Astrophys.* <https://doi.org/10.1051/0004-6361/201833051> (2018).
34. Lindegren, L. et al. Gaia data release 1. Astrometry: one billion positions, two million proper motions and parallaxes. *Astron. Astrophys.* **595**, A4 (2016).
35. Farnocchia, D., Chesley, S. R., Chamberlin, A. B. & Tholen, D. J. Star catalog position and proper motion corrections in asteroid astrometry. *Icarus* **245**, 94–111 (2015).
36. Monet, D. G. et al. The USNO-B catalog. *Astron. J.* **125**, 984–993 (2003).
37. Farnocchia, D. et al. High precision comet trajectory estimates: the Mars flyby of C/2013 A1 (Siding Spring). *Icarus* **266**, 279–287 (2016).
38. Yeomans, D. K. & Chodas, P. W. An asymmetric outgassing model for cometary nongravitational accelerations. *Astron. J.* **98**, 1083–1093 (1989).
39. A'Hearn, M. F. Comets as building blocks. *Annu. Rev. Astron. Astrophys.* **49**, 281–299 (2011).
40. Carry, B. Density of asteroids. *Planet. Space Sci.* **73**, 98–118 (2012).
41. Crovisier, J. & Schloerb, F. P. in *Comets in the Post-Halley Era* (eds Newburn, R. L. et al.) 166 (Kluwer, The Netherlands, 1991).
42. Schorghofer, N. The lifetime of ice on main belt asteroids. *Astrophys. J.* **682**, 697–705 (2008).
43. Laufer, D., Pat-El, I. & Bar-Nun, A. Experimental simulation of the formation of non-circular active depressions on comet Wild-2 and of ice grain ejection from cometary surfaces. *Icarus* **178**, 248–252 (2005).
44. Stern, S. A. ISM-induced erosion and gas-dynamical drag in the Oort cloud. *Icarus* **84**, 447–466 (1990).
45. Prialnik, D. Crystallization, sublimation, and gas release in the interior of a porous comet nucleus. *Astrophys. J.* **388**, 196–202 (1992).
46. Vereš, P., Farnocchia, D., Chesley, S. R. & Chamberlin, A. B. Statistical analysis of astrometric errors for the most productive asteroid surveys. *Icarus* **296**, 139–149 (2017).



Extended Data Fig. 1 | Non-gravitational accelerations of Solar System comets and ‘Oumuamua. Measured non-gravitational radial accelerations A_1 for short-period (red) and long-period (blue) comets from the JPL Small Body Database (<https://ssd.jpl.nasa.gov/sbdb.cgi>). The solid vertical

black line indicates the A_1 value for ‘Oumuamua, which falls within the range observed for Solar System comets; the dashed vertical black lines mark the corresponding 1σ uncertainty.

Extended Data Table 1 | Ground-based astrometry

Date (UTC)	R.A.	Dec.	$\sigma_{\text{R.A.}} (\text{''})$	$\sigma_{\text{Dec.}} (\text{''})$	Obs. code
2017-10-18.472979	01 59 57.460	+02 06 04.02	1.00	1.00	F51
2017-10-18.499898	01 59 08.928	+02 07 20.53	1.50	1.50	F51
2017-10-19.397150	01 34 55.364	+02 45 03.09	0.40	0.40	F51
2017-10-19.408370	01 34 38.761	+02 45 28.19	0.40	0.40	F51
2017-10-19.419685	01 34 21.996	+02 45 53.47	0.40	0.40	F51
2017-10-19.431056	01 34 05.210	+02 46 18.48	1.00	1.00	F51
2017-10-19.940934	01 22 22.290	+03 03 53.82	0.20	0.20	J04
2017-10-19.943901	01 22 18.370	+03 03 59.58	0.20	0.20	J04
2017-10-22.371415	00 40 57.815	+04 02 50.75	0.05	0.05	568
2017-10-22.372590	00 40 56.875	+04 02 52.02	0.05	0.05	568
2017-10-22.373983	00 40 55.762	+04 02 53.49	0.05	0.05	568
2017-10-23.384311	00 28 51.402	+04 19 02.21	0.15	0.15	568
2017-10-23.385548	00 28 50.593	+04 19 03.41	0.15	0.15	568
2017-10-23.386852	00 28 49.730	+04 19 04.54	0.15	0.15	568
2017-10-25.044458	00 13 18.796	+04 39 35.19	0.05	0.05	309
2017-10-25.050182	00 13 15.981	+04 39 38.79	0.05	0.05	309
2017-10-25.061553	00 13 10.389	+04 39 45.94	0.05	0.05	309
2017-10-25.112088	00 12 45.650	+04 40 17.26	0.05	0.05	309
2017-10-25.117597	00 12 42.966	+04 40 20.70	0.05	0.05	309
2017-10-26.133749	00 05 15.166	+04 49 55.54	0.06	0.06	309
2017-10-26.138575	00 05 13.175	+04 49 58.07	0.06	0.06	309
2017-10-26.143286	00 05 11.230	+04 50 00.52	0.06	0.06	309
2017-10-26.185052	00 04 54.100	+04 50 21.91	0.06	0.06	309
2017-10-27.269327	23 58 14.606	+04 58 44.31	0.06	0.06	568
2017-10-27.282873	23 58 09.917	+04 58 50.36	0.06	0.06	568
2017-10-27.304553	23 58 02.427	+04 58 59.94	0.05	0.05	568
2017-10-27.330214	23 57 53.596	+04 59 11.15	0.10	0.10	568
2017-10-27.381822	23 57 35.926	+04 59 33.51	0.10	0.10	568
2017-11-15.306018	23 18 51.738	+06 14 13.51	0.06	0.06	568
2017-11-15.309275	23 18 51.633	+06 14 14.10	0.06	0.06	568
2017-11-15.312534	23 18 51.529	+06 14 14.66	0.06	0.06	568
2017-11-15.315806	23 18 51.418	+06 14 15.25	0.06	0.06	568
2017-11-16.207482	23 18 27.240	+06 16 59.12	0.10	0.10	568
2017-11-16.210740	23 18 27.141	+06 16 59.74	0.10	0.10	568
2017-11-16.213997	23 18 27.045	+06 17 00.34	0.10	0.10	568
2017-11-16.217253	23 18 26.956	+06 17 00.94	0.10	0.10	568
2017-11-21.026940	23 17 05.962	+06 32 01.74	0.10	0.10	304
2017-11-21.032458	23 17 05.893	+06 32 02.84	0.10	0.10	304
2017-11-21.038153	23 17 05.834	+06 32 04.03	0.10	0.10	304
2017-11-21.043922	23 17 05.765	+06 32 05.08	0.10	0.10	304
2017-11-21.060925	23 17 05.573	+06 32 08.23	0.10	0.10	304
2017-11-21.066145	23 17 05.522	+06 32 09.11	0.10	0.10	304
2017-11-21.081650	23 17 05.348	+06 32 12.17	0.10	0.10	304
2017-11-22.222847	23 16 57.168	+06 35 44.32	0.05	0.05	568
2017-11-22.246144	23 16 56.979	+06 35 48.63	0.05	0.05	568
2017-11-22.269437	23 16 56.790	+06 35 53.21	0.05	0.05	568
2017-11-22.292688	23 16 56.602	+06 35 57.52	0.05	0.05	568
2017-11-22.316355	23 16 56.416	+06 36 02.00	0.05	0.05	568
2017-11-23.038940	23 16 53.146	+06 38 25.80	0.12	0.12	304
2017-11-23.070610	23 16 52.967	+06 38 32.02	0.12	0.12	304
2017-11-23.274337	23 16 52.324	+06 39 06.39	0.05	0.05	568
2017-11-23.288299	23 16 52.248	+06 39 09.09	0.10	0.10	568
2017-11-23.373957	23 16 51.831	+06 39 25.49	0.12	0.12	568

Ground-based astrometric positions obtained by our team, with associated 1σ errors, as used in our analysis. For observations with codes F51 or J04, we list the manual re-measurements and associated astrometric errors that we used here, rather than the values available from the Minor Planet Center.

Extended Data Table 2 | HST astrometry

Date (UTC)	R.A.	Dec.	X (km)	Y (km)	Z (km)
2017-11-21.13949584	23:17:05.4011	+06:32:22.611	+1797.7	-6042.7	-2854.2
2017-11-21.14575732	23:17:05.1408	+06:32:24.547	+4946.6	-3541.9	-3298.4
2017-11-21.15201917	23:17:04.8217	+06:32:25.137	+6404.8	+0169.5	-2612.2
2017-11-21.15828066	23:17:04.5335	+06:32:24.580	+5671.7	+3822.4	-1030.1
2017-11-21.16454214	23:17:04.3574	+06:32:23.480	+2994.9	+6164.9	+0905.8
2017-11-21.20571103	23:17:04.8264	+06:32:35.174	+1795.0	-6038.1	-2865.8
2017-11-21.21197288	23:17:04.5677	+06:32:37.090	+4948.6	-3540.6	-3296.8
2017-11-21.21823436	23:17:04.2498	+06:32:37.665	+6410.7	+0166.7	-2598.0
2017-11-21.22449584	23:17:03.9643	+06:32:37.108	+5679.5	+3816.6	-1008.1
2017-11-21.23075732	23:17:03.7895	+06:32:36.014	+3002.0	+6158.1	+0928.0
2017-11-22.53035214	23:16:55.8284	+06:36:47.893	+1959.5	-5866.8	-3104.3
2017-11-22.53661399	23:16:55.5985	+06:36:49.486	+5131.8	-3358.1	-3206.3
2017-11-22.54287547	23:16:55.3158	+06:36:49.816	+6549.9	+0298.0	-2209.5
2017-11-22.54913695	23:16:55.0673	+06:36:49.160	+5726.3	+3851.8	-0454.7
2017-11-22.55539843	23:16:54.9275	+06:36:48.169	+2939.8	+6084.8	+1456.3
2017-12-12.06468176	23:20:53.3768	+07:45:46.658	+1679.8	-6660.5	+0794.3
2017-12-12.07094324	23:20:53.4111	+07:45:47.298	+4666.2	-4478.2	+2443.6
2017-12-12.07720509	23:20:53.3935	+07:45:48.323	+6052.6	-0759.9	+3252.8
2017-12-12.08346657	23:20:53.3819	+07:45:49.844	+5364.0	+3218.8	+2943.9
2017-12-12.08972805	23:20:53.4283	+07:45:51.864	+2836.8	+6094.0	+1623.1
2018-01-02.32061993	23:31:48.3214	+09:16:31.366	+1638.4	-6507.3	+1657.0
2018-01-02.32688178	23:31:48.4836	+09:16:34.181	+4853.0	-4919.9	-0229.1
2018-01-02.33314327	23:31:48.5971	+09:16:36.832	+6402.6	-1644.4	-2036.4
2018-01-02.33940475	23:31:48.7038	+09:16:38.996	+5759.0	+2194.7	-3144.5
2018-01-02.34566623	23:31:48.8412	+09:16:40.537	+3145.7	+5283.4	-3174.2
2018-01-02.45306216	23:31:53.0509	+09:17:08.097	+1635.1	-6519.2	+1612.9
2018-01-02.45932364	23:31:53.2126	+09:17:10.921	+4843.1	-4927.1	-0279.6
2018-01-02.46558512	23:31:53.3287	+09:17:13.563	+6389.8	-1644.4	-2076.1
2018-01-02.47184697	23:31:53.4347	+09:17:15.725	+5747.8	+2202.1	-3159.8
2018-01-02.47810845	23:31:53.5703	+09:17:17.228	+3139.8	+5295.5	-3159.8

Full set of HST-based astrometric positions used here, together with the corresponding geocentric location of the spacecraft in equatorial J2000.0 Cartesian coordinates. Uncertainties of 0.05" were assumed for these observations in our orbital analysis.

Extended Data Table 3 | Uncertainty assumptions for existing astrometry

Obs. code	Date (UTC)	$\sigma_{\text{R.A.}} (\text{''})$	$\sigma_{\text{Dec.}} (\text{''})$
703	2017 October 14, 17	2	2
246	2017 October 19	3	3
Q62	2017 October 22	3	3
G96	2017 October 25	*	*
850	2017 October 27	6	6
H01	2017 October 28, 29, 30	0.3	0.3
705	2017 October 29	3	3
G37	2017 October 30	*	*
H01	2017 November 9, 10, 12	0.3	0.3
G37	2017 November 11	0.3	0.3
H01	2017 November 17	0.5	0.5

Adopted uncertainties for astrometry obtained by other observers and publicly available through the Minor Planet Center. For all observations not listed in this table, we conservatively⁴⁶ adopted uncertainties of 1''. Observations marked with an asterisk in the error columns were deemed unreliable by the respective observers and hence excluded from our analysis. Finally, the uncertainties listed for 703, H01 and G37 were obtained through direct communication with the corresponding observers.

Majorana quantization and half-integer thermal quantum Hall effect in a Kitaev spin liquid

Y. Kasahara¹, T. Ohnishi¹, Y. Mizukami², O. Tanaka², Sixiao Ma¹, K. Sugii³, N. Kurita⁴, H. Tanaka⁴, J. Nasu⁴, Y. Motome⁵, T. Shibauchi² & Y. Matsuda^{1*}

The quantum Hall effect in two-dimensional electron gases involves the flow of topologically protected dissipationless charge currents along the edges of a sample. Integer or fractional electrical conductance is associated with edge currents of electrons or quasiparticles with fractional charges, respectively. It has been predicted that quantum Hall phenomena can also be created by edge currents with a fundamentally different origin: the fractionalization of quantum spins. However, such quantization has not yet been observed. Here we report the observation of this type of quantization of the Hall effect in an insulating two-dimensional quantum magnet¹, α -RuCl₃, with a dominant Kitaev interaction (a bond-dependent Ising-type interaction) on a two-dimensional honeycomb lattice^{2–7}. We find that the application of a magnetic field parallel to the sample destroys long-range magnetic order, leading to a field-induced quantum-spin-liquid ground state with substantial entanglement of local spins^{8–12}. In the low-temperature regime of this state, the two-dimensional thermal Hall conductance reaches a quantum plateau as a function of the applied magnetic field and has a quantization value that is exactly half of the two-dimensional thermal Hall conductance of the integer quantum Hall effect. This half-integer quantization of the thermal Hall conductance in a bulk material is a signature of topologically protected chiral edge currents of charge-neutral Majorana fermions (particles that are their own antiparticles), which have half the degrees of freedom of conventional fermions^{13–16}. These results demonstrate the fractionalization of spins into itinerant Majorana fermions and Z_2 fluxes, which is predicted to occur in Kitaev quantum spin liquids^{1,3}. Above a critical magnetic field, the quantization disappears and the thermal Hall conductance goes to zero rapidly, indicating a topological quantum phase transition between the states with and without chiral Majorana edge modes. Emergent Majorana fermions in a quantum magnet are expected to have a great impact on strongly correlated quantum matter, opening up the possibility of topological quantum computing at relatively high temperatures.

Topological states of matter are described in terms of topological invariant quantities whose values are quantized. The quantity most frequently used to prove the existence of these states is the electrical Hall conductivity. In the quantum Hall state, the Hall conductance σ_{xy}^{2D} is quantized in units of $e^2/2\pi\hbar$, where e is the electronic charge and \hbar is the Planck constant, as $\sigma_{xy}^{2D} = q(e^2/2\pi\hbar)$; q is an integer in the integer quantum Hall effect (QHE) and a fraction in the fractional QHE where, with very few exceptions, it has an odd denominator. These quantizations attest to topologically ordered states. Another topological invariant in the topological phase is the two-dimensional (2D) thermal Hall conductance. The thermal Hall conductivity per 2D sheet, κ_{xy}^{2D} , is quantized in units of $(\pi/6)(k_B^2/\hbar)/T$, where k_B is the Boltzmann constant and T is the temperature, as

$$\kappa_{xy}^{2D}/T = q(\pi/6)(k_B^2/\hbar) \quad (1)$$

Although the thermal Hall conductivity is much harder to measure than electrical Hall conductivity, it has a clear advantage in revealing the topological phases possessing charge-neutral excitations that cannot be detected by the electrical Hall conductivity. In particular, a $q = 1/2$ state with positive thermal Hall sign is a decisive manifestation of the charge-neutral edge currents of Majorana particles (Fig. 1a, b), distinguishing unambiguously between different candidate topological orders. We note that a Majorana quantized phase characterized by $q = 1/2$ has been predicted in chiral topological superconductors^{13–15}. However, as the topological superconductivity in bulk materials has not been fully established, previous experiments searching for Majorana fermions have focused on the proximity effect between conventional superconductors and nanowires or topological materials^{17–20}. Here we present a fundamentally different approach to this issue and perform direct measurements of the thermal Hall conductance in a bulk insulating magnet.

Systems composed of interacting 1/2 spins on a honeycomb lattice with bond-directional exchange interactions J_K are of great interest because they host quantum-spin-liquid (QSL) ground states where topological excitations emerge¹. Such Kitaev QSLs exhibit two types of fractionalized quasiparticle excitation, that is, itinerant (mobile) Majorana fermions and Z_2 fluxes with a gap. The Majorana fermion has a massless (gapless) Dirac-type dispersion in zero field. In magnetic fields, a Majorana fermion system characterized by the bulk gap and gapless edge modes has been realized^{1,3}, and the Z_2 flux obeys anyonic statistics.

Recently, a strongly spin-orbit-coupled Mott insulator, α -RuCl₃, has emerged as a prime candidate for hosting an approximate Kitaev QSL. In this compound, local $j_{\text{eff}} = 1/2$ pseudospins are almost coplanar within the 2D honeycomb layer and the Kitaev interaction, $J_K/k_B \approx 80$ K, has an important role^{5–7}. The system is in a spin-liquid (Kitaev paramagnetic) state below about J_K/k_B and shows antiferromagnetic (AFM) order with zigzag spin structure²¹ (Fig. 1c) at the Néel temperature $T_N \approx 7$ K due to non-Kitaev interactions, such as Heisenberg exchange and off-diagonal interactions. The thermal Hall conductance of α -RuCl₃ has been measured in a magnetic field perpendicular to the 2D planes²². For this geometry, a finite positive κ_{xy}/T emerges in the spin-liquid regime, at $T_N < T \lesssim 80$ K. On entering the AFM state, κ_{xy}/T changes sign and its magnitude is strongly suppressed. The quantization and plateau behaviour of κ_{xy}^{2D}/T have not been observed in the spin-liquid regime. Therefore, expanding the measurements to a lower-temperature region in the liquid state is crucial.

The response of α -RuCl₃ to magnetic fields is highly anisotropic, with largely different in-plane and out-of-plane properties^{8,11,12,23,24}. It has been reported that although T_N is minimally influenced by an external magnetic field perpendicular to the 2D plane, it is markedly suppressed by a parallel field. This highly anisotropic response is confirmed by measurements of the longitudinal thermal conductivity, κ_{xx} , with the heat current along the a axis in a magnetic field H applied along various directions in the a – c plane, as shown in the inset of

¹Department of Physics, Kyoto University, Kyoto, Japan. ²Department of Advanced Materials Science, University of Tokyo, Chiba, Japan. ³Institute for Solid State Physics, University of Tokyo, Chiba, Japan. ⁴Department of Physics, Tokyo Institute of Technology, Tokyo, Japan. ⁵Department of Applied Physics, University of Tokyo, Tokyo, Japan. *e-mail: matsuda@scphys.kyoto-u.ac.jp

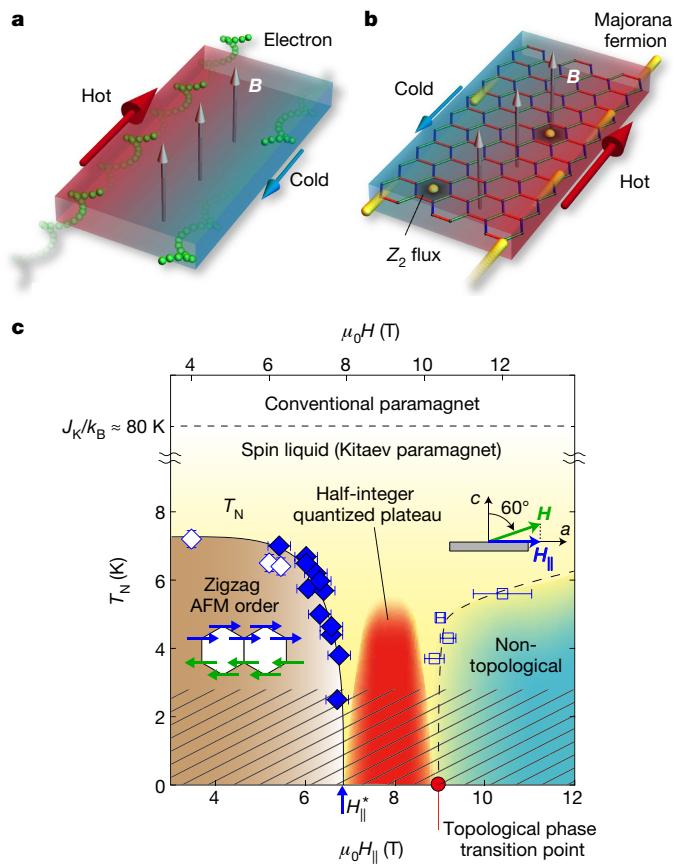


Fig. 1 | Chiral Majorana edge currents and temperature–magnetic field phase diagram of α -RuCl₃. **a, b**, Schematic illustrations of heat conduction in the integer quantum Hall state of a 2D electron gas (**a**) and a Kitaev QSL state (**b**) in a magnetic field perpendicular to the sample plane (grey arrows). In the red (blue) area, the temperature is higher (lower), and the red and blue arrows represent thermal flow. In the quantum Hall state, the skipping orbits of electrons (green spheres) at the edge, which form one-dimensional edge channels, conduct heat and κ_{xy} is negative in sign. In the Kitaev QSL state, spins are fractionalized into Majorana fermions (yellow spheres) and Z_2 fluxes (hexagons). The heat is carried by chiral edge currents of charge-neutral Majorana fermions and κ_{xy} is positive in sign. **c**, Phase diagram of α -RuCl₃ in a field tilted at $\theta = 60^\circ$ (see right inset, where green and blue arrows represent the magnetic field \mathbf{H} and parallel field component H_{\parallel}). Open and closed diamonds represent the onset temperature of AFM order with zigzag-type T_N determined by the T and H dependences of κ_{xx} , respectively (see Fig. 2b and Extended Data Figs. 1 and 2). Below $T \approx J_K/k_B \approx 80$ K, the spin-liquid (Kitaev paramagnetic) state appears. At $\mu_0 H_{\parallel}^* \approx 7$ T, T_N vanishes. A half-integer quantized plateau of the 2D thermal Hall conductance is observed in the red area. Open blue squares represent the fields where the thermal Hall response disappears. The red circle is the suggested topological phase-transition point that separates the non-trivial QSL state with topologically protected chiral Majorana edge currents from a trivial state, such as a non-topological spin liquid. The striped region denotes the region that was not accessible in the thermal Hall effect measurements. Error bars represent one standard deviation (error bars for the temperature are smaller than the symbols). The left inset shows the zigzag magnetic structure in the AFM state. The magnetic moments of Ru atoms represented by blue and green arrows are aligned antiparallel.

Fig. 2a, where $H_{\parallel} = H \sin \theta$ and $H_{\perp} = H \cos \theta$ are the field components parallel and perpendicular to the a axis, respectively, and θ is the angle between \mathbf{H} and the c axis. In zero field, κ_{xx} exhibits a distinct kink at T_N , as shown in Fig. 2a. Although this kink is observed in a perpendicular field ($\theta = 0^\circ$) of 12 T at the same temperature, no such anomaly is observed in a parallel field^{11,12} ($\theta = 90^\circ$) of 7 T. In Fig. 2a, we also plot κ_{xy} in an applied magnetic field of 8 T, tilted away from the c axis ($\theta = 60^\circ$, $\mu_0 H_{\parallel} \approx 7$ T). As in the case of the parallel field, no kink is

observed. Figure 1c displays the phase diagram of an α -RuCl₃ sample in a tilted field of $\theta = 60^\circ$, where T_N is plotted as a function of H_{\parallel} . The inset of Fig. 2b shows T_N plotted as a function of H_{\parallel} for $\theta = 45^\circ$, 60° and 90° . For $\theta = 60^\circ$, T_N agrees well with that for 90° and vanishes at the same critical field of $\mu_0 H_{\parallel}^* \approx 7$ T, whereas for 45° T_N vanishes at $\mu_0 H_{\parallel} \approx 6$ T. Although T_N does not scale perfectly with H_{\parallel} , these results demonstrate the quasi-2D nature of the magnetic properties. In stark contrast to the strong out-of-plane (*a*–*c*) anisotropy, the in-plane (*a*–*b*) anisotropy is very small (Extended Data Fig. 3a–c).

Above $H_{\parallel} = H_{\parallel}^*$, where the AFM order melts, the presence of a peculiar spin-liquid state has been suggested on the basis of nuclear magnetic resonance and neutron scattering measurements; the former show the presence of a spin gap²⁵ and the latter reveal unusual continuous spin excitations²⁶. These magnetic properties are consistent with those expected in a Kitaev-type spin-liquid state.

To study the thermal Hall effect in the spin-liquid state above $H_{\parallel} = H_{\parallel}^*$, κ_{xy} is measured by sweeping fields in tilted directions and obtained by anti-symmetrizing the thermal response of the sample with respect to the field direction. In this configuration, the Hall response is determined by H_{\perp} . Because the magnitude of κ_{xy} is extremely small compared to κ_{xx} in α -RuCl₃, special care is taken to detect the intrinsic thermal Hall signal (see Methods). Figure 3a–d and Fig. 3e–h depict κ_{xy}/T at $\theta = 60^\circ$ and 45° , respectively, plotted as a function of H_{\perp} above $H_{\parallel} = H_{\parallel}^*$ at low temperatures. The experimental error in the detection of the temperature difference between Hall contacts becomes considerable below 3.5 K, leading to unreliable determination of κ_{xy} in our setup.

In the AFM state, κ_{xy}/T is extremely small (see Extended Data Fig. 4). Upon entering the field-induced spin-liquid state, κ_{xy}/T , which is positive in sign, increases rapidly. The most striking feature is that κ_{xy}/T exhibits a plateau in the field range of 4.5 T $< \mu_0 H_{\perp} < 4.8$ – 5.0 T for $\theta = 60^\circ$ and 6.8 T $< \mu_0 H_{\perp} < 7.2$ – 7.4 T for $\theta = 45^\circ$, as shown in Fig. 3a–c and Fig. 3e–g, respectively. The right axes represent κ_{xy}^{2D}/T in units of quantum thermal Hall conductance ($\pi/6(k_F^2\hbar)$, where $\kappa_{xy}^{2D} = \kappa_{xy}d$ with a layer distance²¹ of $d = 5.72$ Å. Remarkably, the plateau is very close to the half of the quantum thermal Hall conductance reported in the integer quantum Hall system²⁷ within the error of 3%, demonstrating the emergence of a half-integer thermal Hall conductance plateau. Above $\mu_0 H_{\perp} \approx 5.0$ T for $\theta = 60^\circ$ (7.4 T for $\theta = 45^\circ$), κ_{xy}^{2D}/T decreases rapidly and vanishes. We note that the half-integer quantized plateau is reproduced in crystal from different growth (Extended Data Fig. 5). Although the plateau behaviour seems to be preserved at 5.6 K, κ_{xy}^{2D}/T slightly deviates from the quantized value. At higher temperatures, the plateau behaviour disappears (Fig. 3d, h).

The temperature dependence of κ_{xy}/T at magnetic fields where a plateau is observed is shown in Fig. 4. The half-integer thermal Hall conductance is observable up to about 5.5 K, above which κ_{xy}/T increases rapidly with T . As shown in the inset of Fig. 4, κ_{xy}/T decreases after reaching a maximum at around 15 K and nearly vanishes above about 60 K (see Extended Data Fig. 6). As the vanishing temperature of κ_{xy}/T is close to the Kitaev interaction, it is natural to consider that the finite thermal Hall signal reflects unusual quasiparticle excitations inherent to the spin-liquid state governed by the Kitaev interaction (see Methods for further discussion).

In equation (1), the coefficient q gives the chiral central charge of the gapless boundary modes, which propagate along one direction. The central charge represents a degree of freedom of one-dimensional gapless modes; it is unity for conventional fermions and $1/2$ for Majorana fermions whose degrees of freedom are half of those of conventional fermions. An integer quantum Hall system with bulk Chern number ν has ν boundary modes with $q = \nu$, whereas a Kitaev QSL with Chern number ν has ν Majorana boundary modes with $q = \nu/2$. Thus, the observed half-integer thermal Hall conductance provides direct evidence of chiral Majorana edge currents. We also note that the positive Hall sign is also consistent with that predicted in the Kitaev QSL¹. In the pure Kitaev model, the excitation energy of the Z_2 flux is estimated⁷ to be $\Delta_F/k_B \approx 0.06 J_K/k_B \approx 5.5$ K. Recent numerical results¹⁶ of the thermal

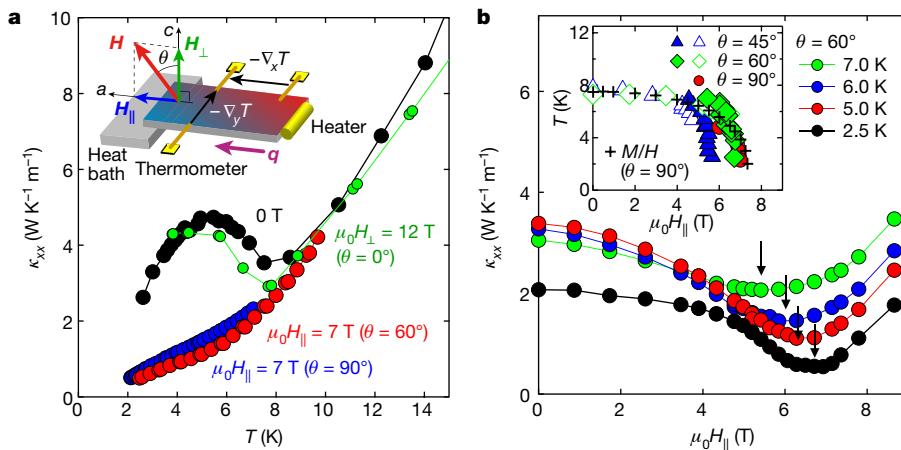


Fig. 2 | Longitudinal thermal conductivity in α -RuCl₃. **a**, Temperature dependence of κ_{xx} in a magnetic field \mathbf{H} applied along various directions in the a - c plane. The inset illustrates a schematic of the measurement setup for κ_{xx} and κ_{xy} (see Methods for details). **b**, κ_{xx} at $\theta = 60^\circ$, plotted as a function of the parallel field component, H_{\parallel} . The inset shows T_N versus H_{\parallel}

at different field directions. T_N is determined by the T dependence of κ_{xx} shown in **a** (open symbols) and by the minimum in the H dependence of κ_{xx} (filled symbols), shown by arrows in the main panel. Crosses show T_N for $\theta = 90^\circ$, determined from magnetic susceptibility (M/H , where M is the magnetization) measurements²⁶.

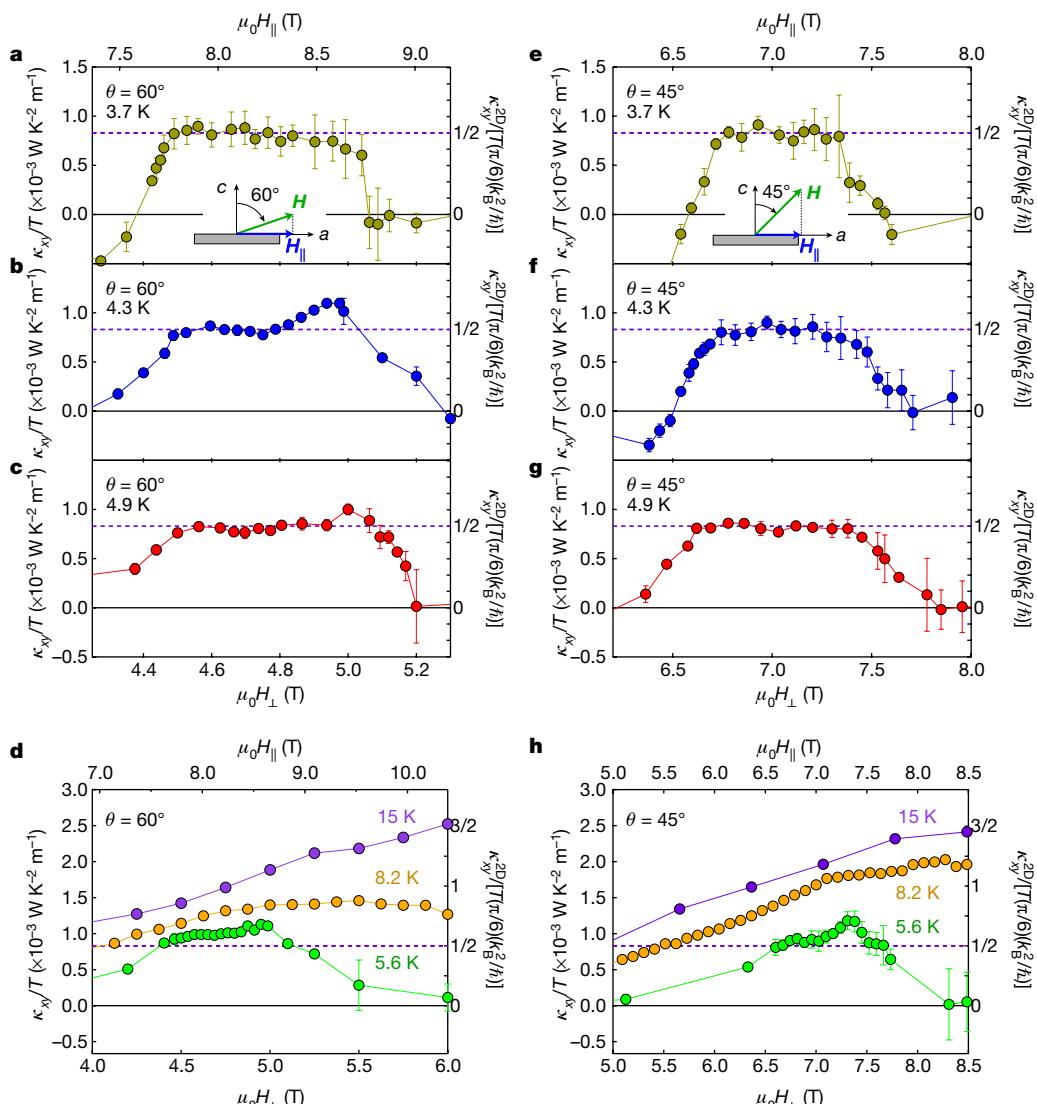


Fig. 3 | Half-integer thermal Hall conductance plateau. **a–h**, Thermal Hall conductivity κ_{xy}/T in a field tilted at $\theta = 60^\circ$ (**a–d**) and 45° (**e–h**) plotted as a function of H_{\perp} (see inset of Fig. 2a). The top axes show the parallel field component, H_{\parallel} . The right scales represent the 2D thermal

Hall conductance, κ_{xy}^{2D}/T , in units of $(\pi/6)(k_B^2/h)$. Violet dashed lines represent the half-integer thermal Hall conductance, $\kappa_{xy}^{2D}/[T(\pi/6)(k_B^2/h)] = 1/2$. Error bars represent one standard deviation.

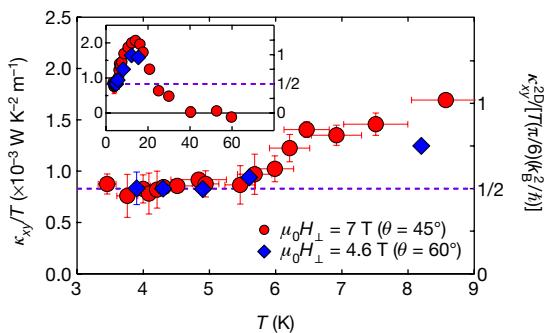


Fig. 4 | Temperature dependence of the thermal Hall conductance.

The main panel shows κ_{xy}/T in fields tilted at $\theta = 45^\circ$ and 60° (see inset of Fig. 2a) at $\mu_0 H_{\perp} = 7$ T and 4.6 T, respectively, where a quantized thermal Hall conductance plateau is observed at low temperatures. The right scale is the 2D thermal Hall conductance κ_{xy}^{2D}/T in units of $(\pi/6)(k_B^2/h)$. The violet dashed line represents the half-integer thermal Hall conductance, $\kappa_{xy}^{2D}/[T(\pi/6)(k_B^2/h)] = 1/2$. The inset shows the same data in a wider temperature regime. Error bars represent one standard deviation.

Hall conductance for the 2D pure Kitaev model calculated with the quantum Monte Carlo method show that quantization occurs slightly below Δ_F/k_B . Experimentally, Δ_F/k_B is estimated²⁵ to be 10 K, which is consistent with the persistence of the thermal Hall quantization up to around 5 K.

In the plateau regime of κ_{xy} , no anomaly is observed in κ_{xx} , probably because phonon contributions largely dominate over fermionic excitations arising from spins in κ_{xx} in the whole temperature range^{28,29}. Moreover, owing to the strong spin-phonon coupling in α -RuCl₃¹¹, the phonon conductivity is expected to show complicated H and T dependences. The observed behaviour of the plateau as a function of H and T therefore demonstrates that κ_{xy}/T is not affected by spin-phonon scattering in the plateau regime, providing strong support for topological protection. The fact that κ_{xy} vanishes at the highest fields, as shown in Fig. 3a–c, e–g, provides direct evidence that the thermal Hall effect is not influenced by phonons, demonstrating that κ_{xy} is a unique and powerful probe in the search for Majorana quantization.

We stress that a half-integer thermal Hall conductance in a bulk material is a direct consequence of the chiral Majorana edge current. Recent experiments based on the proximity effect between a quantum anomalous Hall insulator and a conventional superconductor have reported a signature of chiral Majorana edge modes²⁰. However, this is based on the observation of half-integer quantization of the longitudinal electrical conductance via the scattering matrix effect between the edge states of the insulator and superconductor. Moreover, Majorana fermions in Kitaev magnets and topological superconductors have essentially different features. In the former, strong correlations give rise to Majorana fermions, whereas in the latter they do not play a role. In addition, Majorana fermions exist inside the bulk of a sample in the Kitaev QSL state, in sharp contrast to topological superconductors, where they appear only at the edges. This distinct nature of Majorana fermions is supported by the fact that the quantum plateau disappears below about 400 mK in a topological superconductor device²⁰, whereas it is preserved up to around 5 K in α -RuCl₃.

At $\theta = 60^\circ$, $\kappa_{xy}^{2D}(H)/T$ increases slightly from the quantized value before going to zero at a high field at 4.3 K and 4.9 K, which is reproduced in a different crystal (Extended Data Fig. 5a). However, such a behaviour is not observed at $\theta = 45^\circ$. On the other hand, an overshoot is also observed in the temperature dependence of κ_{xy}^{2D} , irrespective of the angle (Fig. 4) and crystal (Extended Data Fig. 5b); therefore, there seem to be certain high-energy corrections that are responsible for the excess conductivity at high fields and high temperatures. These overshoots are in contrast to the numerical results of the thermal Hall effect for the 2D pure Kitaev model with a weak magnetic field¹⁶. Meanwhile, it has been pointed out that non-Kitaev interactions, such as Heisenberg and off-diagonal ones, are important for α -RuCl₃^{30,31}. Hence, the

discrepancy may be attributed to high-field effects or non-Kitaev interactions, which deserves further study.

The near vanishing of κ_{xy}^{2D}/T after its rapid suppression in the high-field regime (Fig. 3a–c, e–g) demonstrates the disappearance of chiral Majorana edge currents. As shown by the open blue square in Fig. 1c, the temperature at which κ_{xy}^{2D}/T vanishes decreases rapidly with decreasing H_{\parallel} . This suggests a topological quantum phase transition from the non-trivial QSL to a trivial high-field state, where the thermal Hall effect is absent, at $\mu_0 H_{\parallel} \approx 9$ T, as shown by the red circle in Fig. 1c³². The specific heat at 0.47 K for $\theta = 60^\circ$ exhibits a dip-like anomaly in the vicinity of 9 T, which can be associated with an abrupt change of the spin gap at the topological transition, strongly supporting the presence of a characteristic field revealed by κ_{xy}/T (Extended Data Fig. 7a–c). The vanishing of κ_{xy}/T at the highest fields is unlikely to be due to the crossover to a simple forced ferromagnetic state because the magnetization at 9 T is less than 1/3 of the fully polarized value, indicating that paramagnetic spins still remain. The observation of half-integer thermal Hall conductance reveals that topologically protected chiral Majorana edge currents persist in α -RuCl₃, even in the presence of non-Kitaev interactions and a parallel field. This observation opens a possibility of using Majorana fermions and their link to non-Abelian anyons, which are important for topological quantum computing, revealing novel aspects of strongly correlated topological quantum matters.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0274-0>.

Received: 13 November 2017; Accepted: 24 April 2018;

Published online 11 July 2018.

1. Kitaev, A. Anyons in an exactly solved model and beyond. *Ann. Phys.* **321**, 2–111 (2006).
2. Jackeli, G. & Khaliullin, G. Mott insulators in the strong spin-orbit coupling limit: from Heisenberg to a quantum compass and Kitaev models. *Phys. Rev. Lett.* **102**, 017205 (2009).
3. Trebst, S. Kitaev materials. Preprint at <https://arxiv.org/abs/1701.07056> (2017).
4. Kim, H.-S., Shankar, V. V., Catuneanu, A. & Kee, H.-Y. Kitaev magnetism in honeycomb RuCl₃ with intermediate spin-orbit coupling. *Phys. Rev. B* **91**, 241110 (2015).
5. Banerjee, A. et al. Proximate Kitaev quantum spin liquid behaviour in a honeycomb magnet. *Nat. Mater.* **15**, 733–740 (2016).
6. Sandilands, L. J., Tian, Y., Plumb, W., Kim, Y.-J. & Burch, K. S. scattering continuum and possible fractionalized excitations in α -RuCl₃. *Phys. Rev. Lett.* **114**, 147201 (2015).
7. Nasu, J., Knolle, J., Kovrizhin, D. L., Motome, Y. & Moessner, R. Fermionic response from fractionalization in an insulating two-dimensional magnet. *Nat. Phys.* **12**, 912–915 (2016).
8. Yadav, R. et al. Kitaev exchange and field-induced quantum spin-liquid states in honeycomb α -RuCl₃. *Sci. Rep.* **6**, 37925 (2016).
9. Baek, S.-H. et al. Evidence for a field-induced quantum spin liquid in α -RuCl₃. *Phys. Rev. Lett.* **119**, 037201 (2017).
10. Wolter, A. U. B. et al. Field-induced quantum criticality in the Kitaev system α -RuCl₃. *Phys. Rev. B* **96**, 041405 (2017).
11. Leahy, I. A. et al. Anomalous thermal conductivity and magnetic torque response in the honeycomb magnet α -RuCl₃. *Phys. Rev. Lett.* **118**, 187203 (2017).
12. Henrich, R. et al. Unusual phonon heat transport in α -RuCl₃: strong spin-phonon scattering and field-induced spin gap. *Phys. Rev. Lett.* **120**, 117204 (2018).
13. Read, N. & Green, D. Paired states of fermions in two dimensions with breaking of parity and time-reversal symmetries and the fractional quantum Hall effect. *Phys. Rev. B* **61**, 10267–10297 (2000).
14. Sumiyoshi, H. & Fujimoto, S. Quantum thermal hall effect in a time-reversal-symmetry- broken topological superconductor in two dimensions: approach from bulk calculations. *J. Phys. Soc. Jpn.* **82**, 023602 (2013).
15. Nomura, K., Ryu, S., Furusaki, A. & Nagaosa, N. Cross-correlated responses of topological superconductors and superfluids. *Phys. Rev. Lett.* **108**, 026802 (2012).
16. Nasu, J., Yoshitake, J. & Motome, Y. Thermal transport in the Kitaev model. *Phys. Rev. Lett.* **119**, 127204 (2017).
17. Mourik, V. et al. Signatures of Majorana fermions in hybrid superconductor–semiconductor nanowire devices. *Science* **336**, 1003–1007 (2012).

18. Nadj-Perge, S. et al. Observation of Majorana fermions in ferromagnetic atomic chains on a superconductor. *Science* **346**, 602–607 (2014).

19. Das, A. et al. Zero-bias peaks and splitting in an Al-InAs nanowire topological superconductor as a signature of Majorana fermions. *Nat. Phys.* **8**, 887–895 (2012).

20. He, Q. L. et al. Chiral Majorana fermion modes in a quantum anomalous Hall insulator–superconductor structure. *Science* **357**, 294–299 (2017).

21. Johnson, R. D. et al. Monoclinic crystal structure of α -RuCl₃ and the zigzag antiferromagnetic ground state. *Phys. Rev. B* **92**, 235119 (2015).

22. Kasahara, Y. et al. Unusual thermal Hall effect in a Kitaev spin liquid candidate α -RuCl₃. *Phys. Rev. Lett.* **120**, 217205 (2018).

23. Majumder, M., Schmidt, M., Rosner, H., Tsirlin, A. A., Yasuoka, H. & Baenitz, M. Anisotropic Ru³⁺ 4d⁵ magnetism in the α -RuCl₃ honeycomb system: susceptibility, specific heat, and zero-field NMR. *Phys. Rev. B* **91**, 180401 (2015).

24. Chaloupka, L. & Khalilullin, G. Magnetic anisotropy in the Kitaev model systems Na₂IrO₃ and RuCl₃. *Phys. Rev. B* **94**, 064435 (2016).

25. Janša N. et al. Observation of two types of fractional excitation in the Kitaev honeycomb magnet. *Nat. Phys.* <https://doi.org/10.1038/s41567-018-0129-5> (2018).

26. Banerjee, A. et al. Excitations in the field-induced quantum spin liquid state of α -RuCl₃. *npj Quantum Mater.* **3**, 8 (2018).

27. Banerjee, M. et al. Observed quantization of anyonic heat flow. *Nature* **545**, 75–79 (2017).

28. Hirobe, D., Sato, M., Shiomi, Y., Tanaka, H. & Saitoh, E. Magnetic thermal conductivity far above the Néel temperatures in the Kitaev-magnet candidate α -RuCl₃. *Phys. Rev. B* **95**, 241112 (2017).

29. Yu, Y. J. et al. Ultralow-temperature thermal conductivity of the Kitaev honeycomb magnet α -RuCl₃ across the field-induced phase transition. *Phys. Rev. Lett.* **120**, 067202 (2018).

30. Gohlke, M., Wachtel, G., Yamaji, Y., Pollmann, F. & Kim, Y. B. Signatures of quantum spin liquid in Kitaev-like frustrated magnets. *Phys. Rev. B* **97**, 075126 (2018).

31. Winter, S. M., Li, Y., Jeschke, H. O. & Valenti, R. Challenges in design of Kitaev materials: magnetic interactions from competing energy scales. *Phys. Rev. B* **93**, 214431 (2016).

32. Jiang, H.-C., Gu, Z.-C., Qi, X.-L. & Trebst, S. Possible proximity of the Mott insulating iridate Na₂IrO₃ to a topological phase: Phase diagram of the Heisenberg-Kitaev model in a magnetic field. *Phys. Rev. B* **83**, 245104 (2011).

Acknowledgements We thank S. Fujimoto, H. Ishizuka, N. Kawakami, H.-Y. Kee, Y. B. Kim, E.-G. Moon, N. P. Ong, M. Shimoza, M. Udagawa and M. Yamashita for useful discussions. We thank N. Abe, Y. Tokunaga and T. Arima for support in X-ray diffraction measurements. This work was supported by Grants-in-Aid for Scientific Research (KAKENHI) (numbers 25220710, 15H02014, 15H02106, 15H05457, 15K13533, 15K17692, 16H02206, 16H00987, 16K05414, 17H01142 and 18H04223) and Grants-in-Aid for Scientific Research on innovative areas “Topological Materials Science” (number JP15H05852) from Japan Society for the Promotion of Science (JSPS).

Reviewer information *Nature* thanks K.-Y. Choi, K. Shtengel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.K. and Y. Matsuda conceived and designed the study. Y.K., T.O. and S.M. performed the thermal transport measurements. Y. Mizukami, O.T. and K.S. performed the specific heat measurements. N.K. and H.T. synthesized the high-quality single crystalline samples. Y.K., T.O., J.N., Y. Motome, T.S. and Y. Matsuda discussed the results. Y.K., J.N., Y. Motome, T.S. and Y. Matsuda prepared the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0274-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Y.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Single-crystal growth. High-quality single crystals of α -RuCl₃ were grown by a vertical Bridgman method as described in ref.³³. For thermal transport measurements, we carefully picked up thin crystals with a plate-like shape. Typical sample size was roughly 2 mm \times 0.5 mm \times 0.02 mm. We selected the best crystals, in which no anomaly associated with the magnetic transition at 14 K due to the stacking faults was detected by magnetic susceptibility, specific heat and thermal transport measurements.

Thermal transport measurements. Thermal and thermal Hall conductivities were measured simultaneously on the same crystal by the standard steady-state method, using the experimental setup illustrated in the inset of Fig. 2a. A heat current \mathbf{q} was applied along the a axis ($\mathbf{q} \parallel \mathbf{x}$). Using special jigs, a magnetic field \mathbf{H} was applied along various directions in the a - c plane within an accuracy of less than one degree. The temperature gradient $-\nabla_x T \parallel \mathbf{x}$ and $-\nabla_y T \parallel \mathbf{y}$ was measured by carefully calibrated Cernox thermometers. The sample temperature was measured with an accuracy of 0.1 mK using alternating current resistance bridges. A 1-k Ω chip resistor was used to generate the heat current. The magnitude of the thermal gradient was less than 5% of the base temperature. To reduce the noise level, all measurements were performed in a radio-frequency-shielded room. For the measurements of the thermal Hall effect, we removed the longitudinal response due to misalignment of the contacts by anti-symmetrizing the measured $\nabla_y T$ as $\nabla_y T^{\text{asym}}(H) = [\nabla_y T(H) - \nabla_y T(-H)]/2$ at each temperature. We note that the offset transverse thermal gradient due to the misalignment of the Hall contact was reduced to be less than 0.5% of the longitudinal thermal gradient in zero field. κ_{xx} and κ_{xy} were obtained from the longitudinal thermal resistivity, $w_{xx} = \nabla_x T/q$, and the thermal Hall resistivity, $w_{xy} = \nabla_y T^{\text{asym}}/q$, as $\kappa_{xx} = w_{xx}/(w_{xx}^2 + w_{xy}^2)$ and $\kappa_{xy} = w_{xy}/(w_{xx}^2 + w_{xy}^2)$. To avoid a background Hall signal, a LiF heat bath and non-metallic grease were used. We confirmed that the thermal Hall signal in LiF is negligibly small within our experimental resolution³⁴. The experimental error in determining κ_{xy} , caused by the uncertainty in measuring the distance between the contacts and the thickness of the crystal, is within $\pm 2\%$.

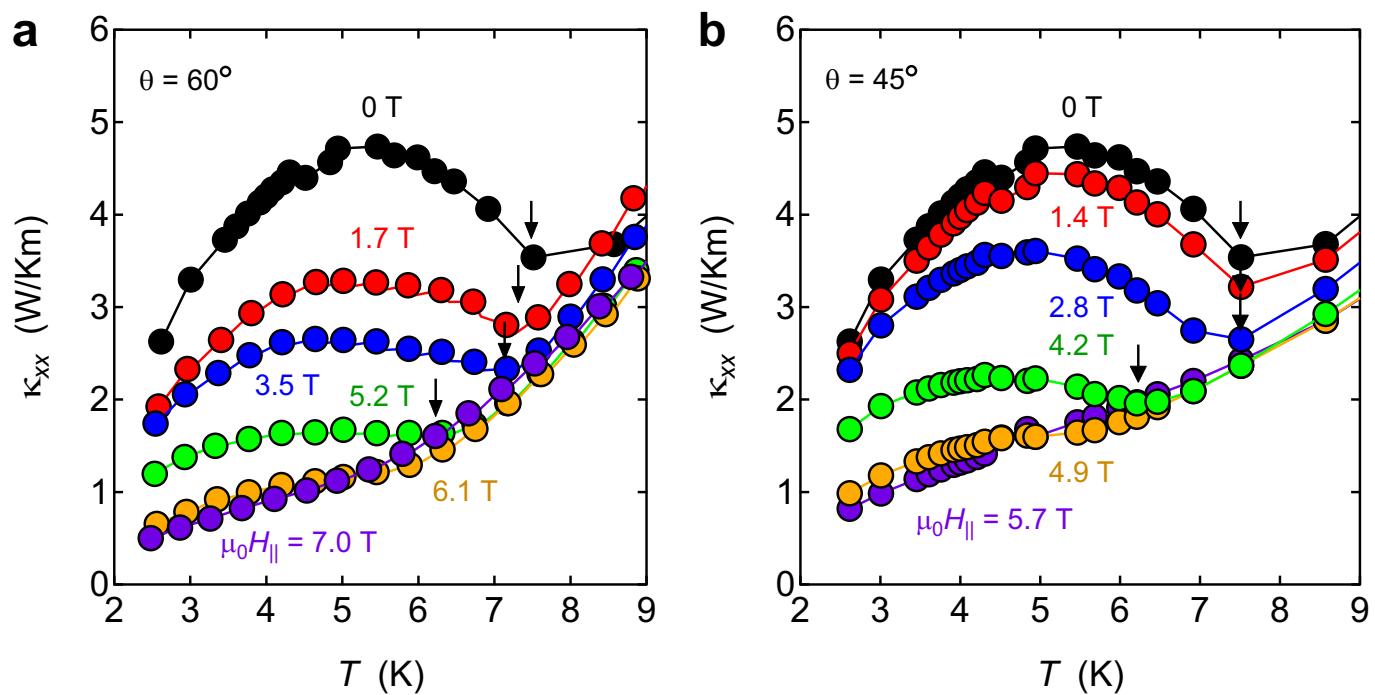
Specific heat measurements. Specific heat was measured by a long relaxation method³⁵ in a ³He cryostat. A Cernox chip resistor was used as both

a thermometer and a heater. The sample was attached to the calorimeter using grease. The thermometer was calibrated in magnetic field of up to 12 T.

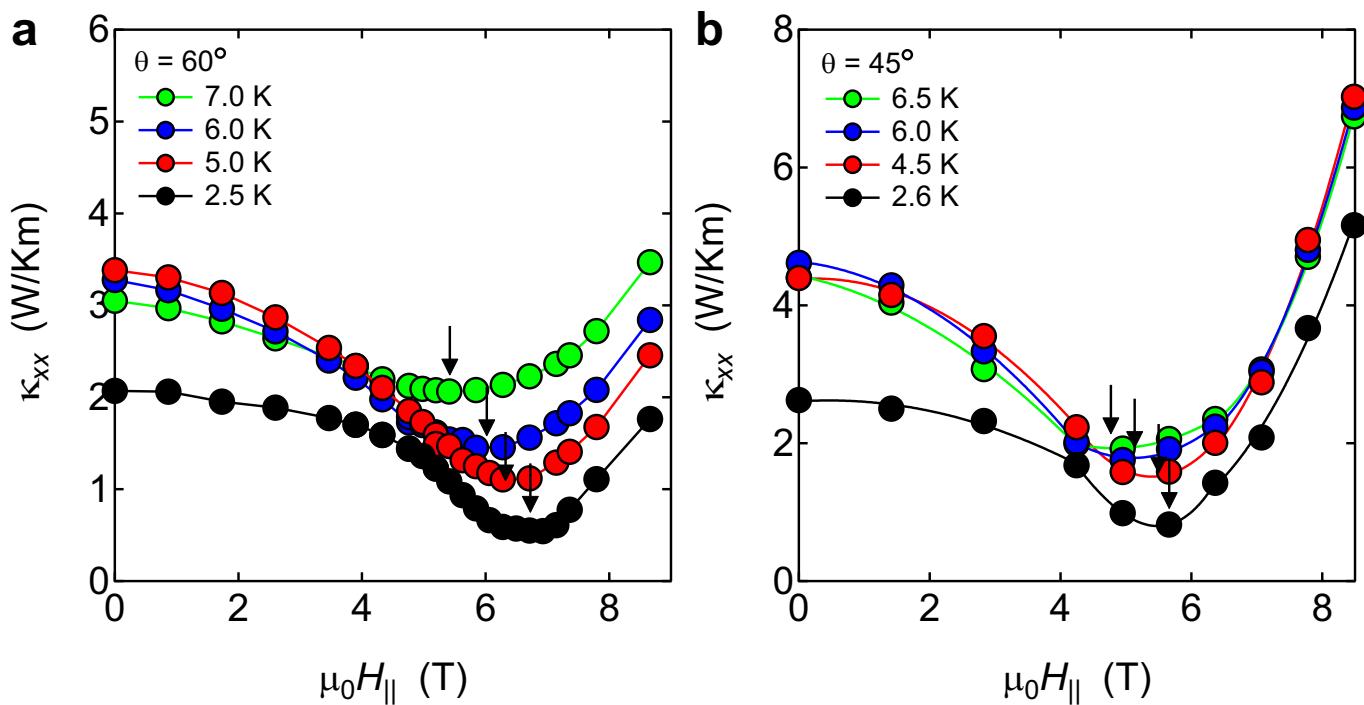
Origin of thermal Hall response. Here we discuss κ_{xy}/T in the high-temperature spin-liquid regime, where no plateau behaviour is observed. A finite κ_{xy}/T in the spin-liquid states has been reported only in the kagomé insulator volborthite Cu₃V₂O₇(OH)₂·2H₂O so far³⁴. We point out that the behaviour of κ_{xy}/T in the high-temperature regime of α -RuCl₃ is essentially different from that in the liquid state of volborthite; the κ_{xy} value of volborthite is opposite in sign to that of α -RuCl₃ and its magnitude is more than one order magnitude smaller. Until now, all theories except the Kitaev model predict that a finite κ_{xy} can appear in spin-liquid states when the Dzyaloshinsky–Moriya (DM) interaction is present³⁶. In fact, volborthite has a large DM interaction. However, the DM interaction in α -RuCl₃ is approximately 5 K, which is much smaller³¹ than J_K , and hence it does not play an important role at high temperatures. Moreover, the phonon thermal Hall conductivity is three orders of magnitude smaller than the observed κ_{xy}/T in the spin-liquid state and shows essentially different temperature dependence³⁷.

Data availability. The data that support the results presented in this paper and other findings of this study are available from the corresponding author upon reasonable request.

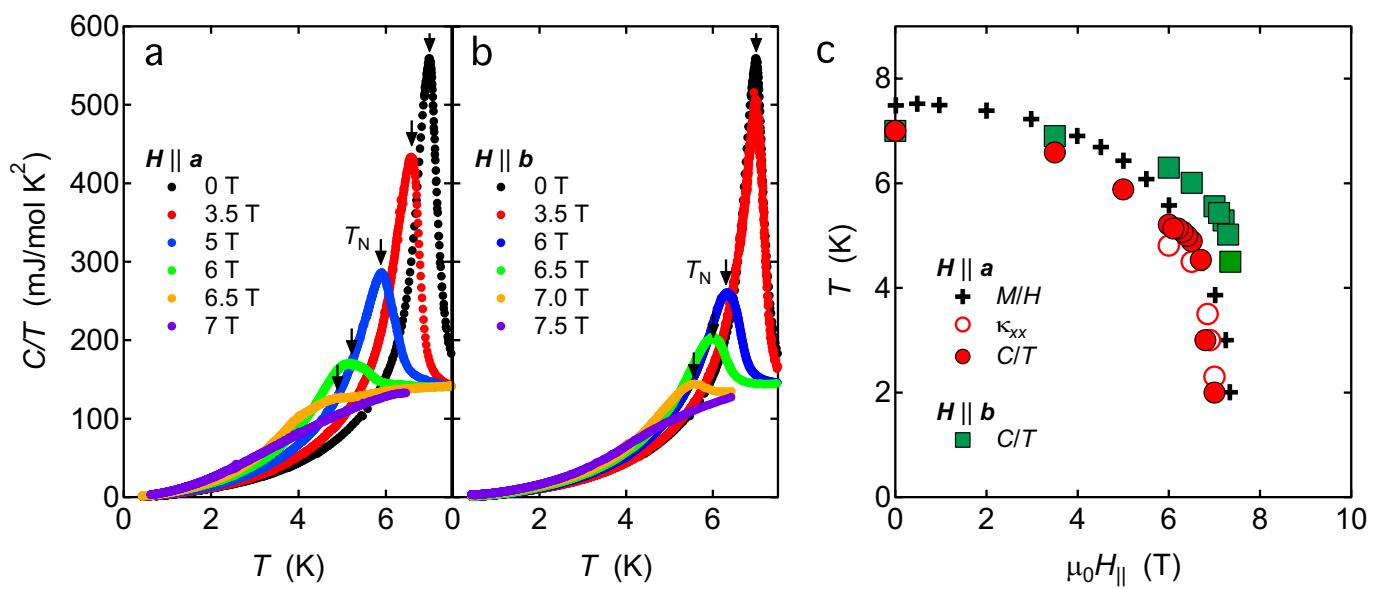
33. Kubota, Y., Tanaka, H., Ono, T., Narumi, Y. & Kindo, K. Successive magnetic phase transition in α -RuCl₃: XY-like frustrated magnet on the honeycomb lattice. *Phys. Rev. B* **91**, 094422 (2015).
34. Watanabe, D. et al. Emergence of nontrivial magnetic excitations in a spin liquid state of kagomé volborthite. *Proc. Natl Acad. Sci. USA* **113**, 8653–8657 (2016).
35. Taylor, O. J., Carrington, A. & Schlueter, J. A. Specific-heat measurements of the gap structure of the organic superconductor κ -(ET)₂Cu[N(CN)₂]Br and κ -(ET)₂Cu(NCS)₂. *Phys. Rev. Lett.* **99**, 057001 (2007).
36. Han, J. H. & Lee, H. Spin chirality and Hall-like transport phenomena of spin excitations. *J. Phys. Soc. Jpn* **86**, 011007 (2017).
37. Sugii, K. et al. Thermal Hall effect in a phonon-glass Ba₃CuSb₂O₉. *Phys. Rev. Lett.* **118**, 145902 (2017).



Extended Data Fig. 1 | Temperature dependence of the longitudinal thermal conductivity. **a, b**, κ_{xx} in a field tilted at $\theta = 60^\circ$ (**a**) and 45° (**b**), plotted as a function of temperature (see inset of Fig. 2a). Arrows indicate the onset temperature of the AFM order T_N .

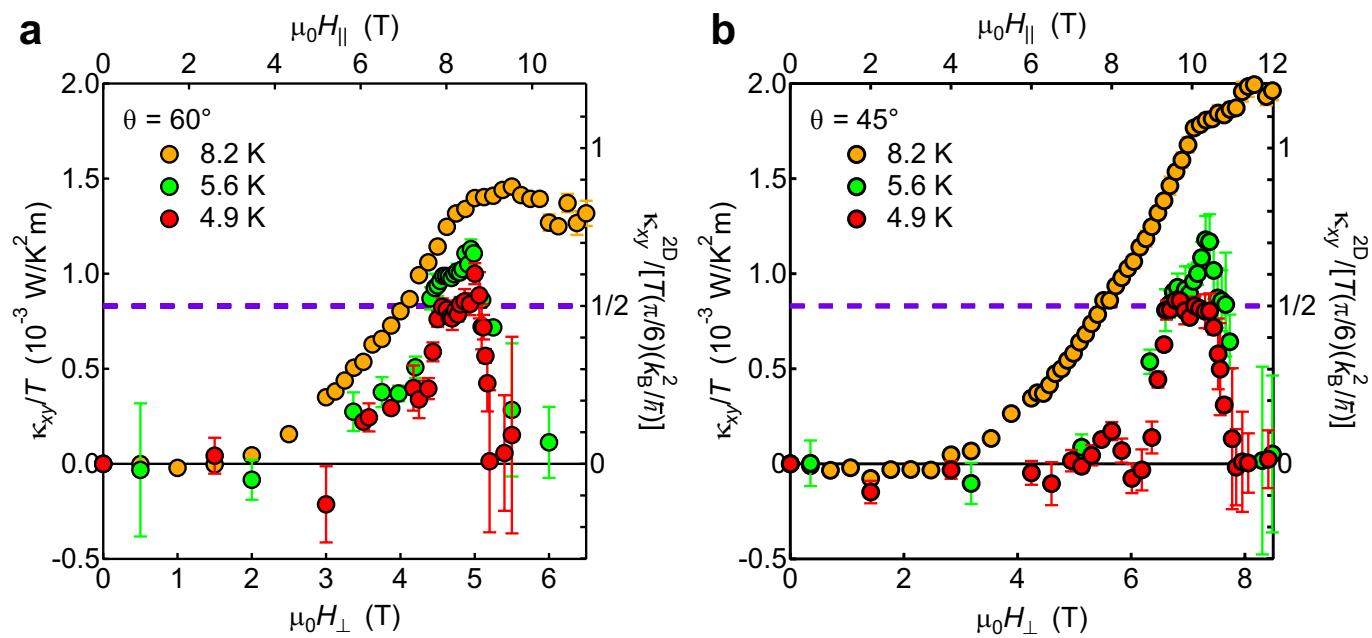


Extended Data Fig. 2 | Field dependence of the longitudinal thermal conductivity. a, b, κ_{xx} in field tilted at $\theta = 60^\circ$ (a) and 45° (b), plotted as a function of the parallel field component H_{\parallel} (see inset of Fig. 2a). Arrows indicate the minimum of κ_{xx} , which is attributed to the onset field of the AFM order.



Extended Data Fig. 3 | Phase diagram of α -RuCl₃ for $H \parallel a$ and $H \parallel b$.
a, b, Temperature dependence of the specific heat, C , divided by T for $H \parallel a$ (a) and $H \parallel b$ (b). Arrows indicate the Néel temperature T_N .
c, Field dependence of T_N for $H \parallel a$ and $H \parallel b$, determined by the specific

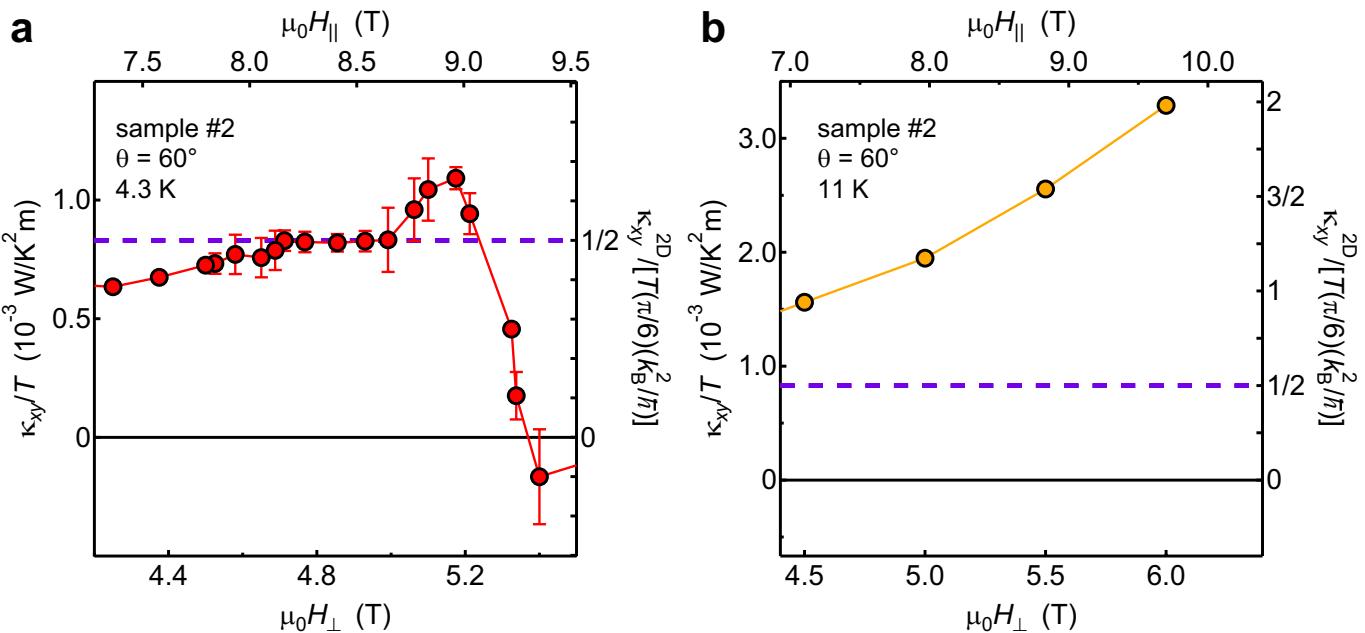
heat measurements. T_N , determined from the thermal conductivity and magnetic susceptibility²⁶, is also shown. The critical field for $H \parallel a$ is slightly lower than that for $H \parallel b$, but both phase diagrams are very similar.



Extended Data Fig. 4 | Field dependence of thermal Hall conductivity.

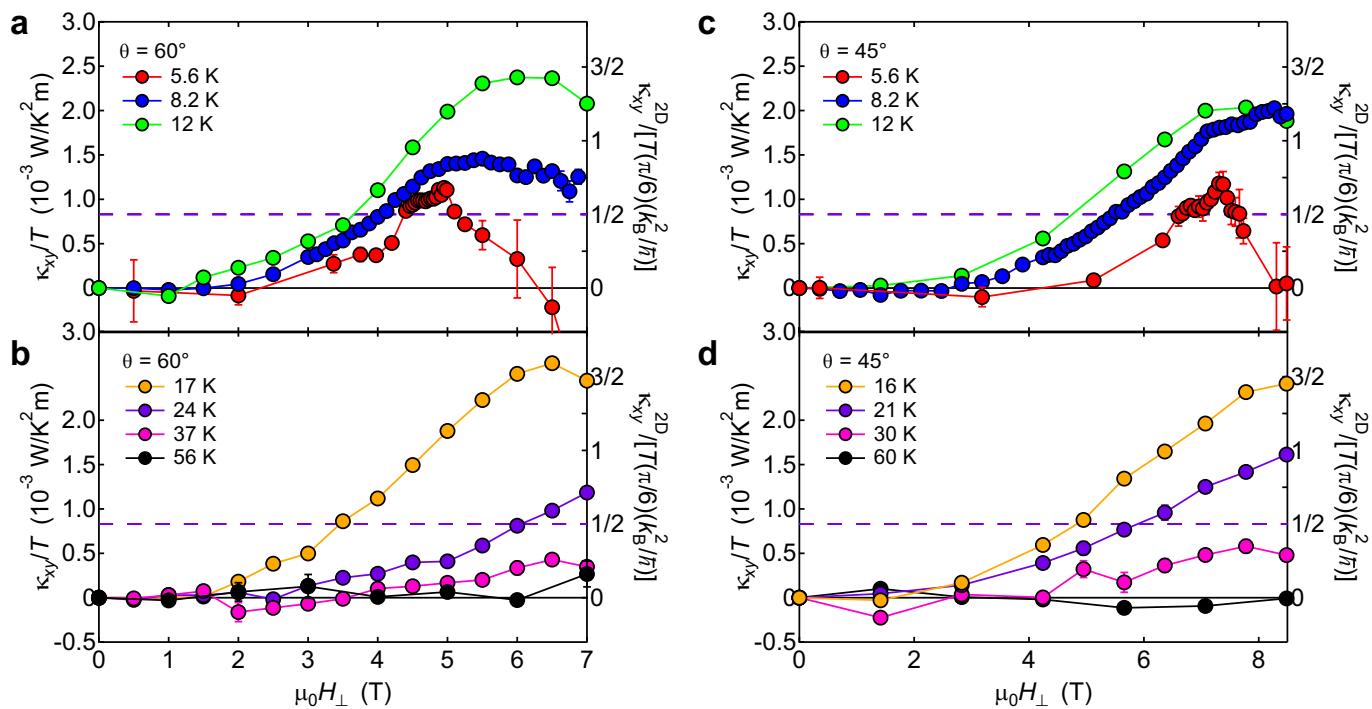
a, b, Thermal Hall conductivity, κ_{xy}/T , in a field tilted at $\theta = 60^\circ$ (a) and 45° (b), plotted as a function of H_\perp (see inset of Fig. 2a). The top axes show the parallel field component, H_\parallel . The right scales represent the 2D thermal

Hall conductance, κ_{xy}^{2D}/T , in units of $(\pi/6)(k_B^2/\hbar)$. Violet dashed lines represent the half-integer thermal Hall conductance, $\kappa_{xy}^{2D}/[T(\pi/6)(k_B^2/\hbar)] = 1/2$. Error bars represent one standard deviation.



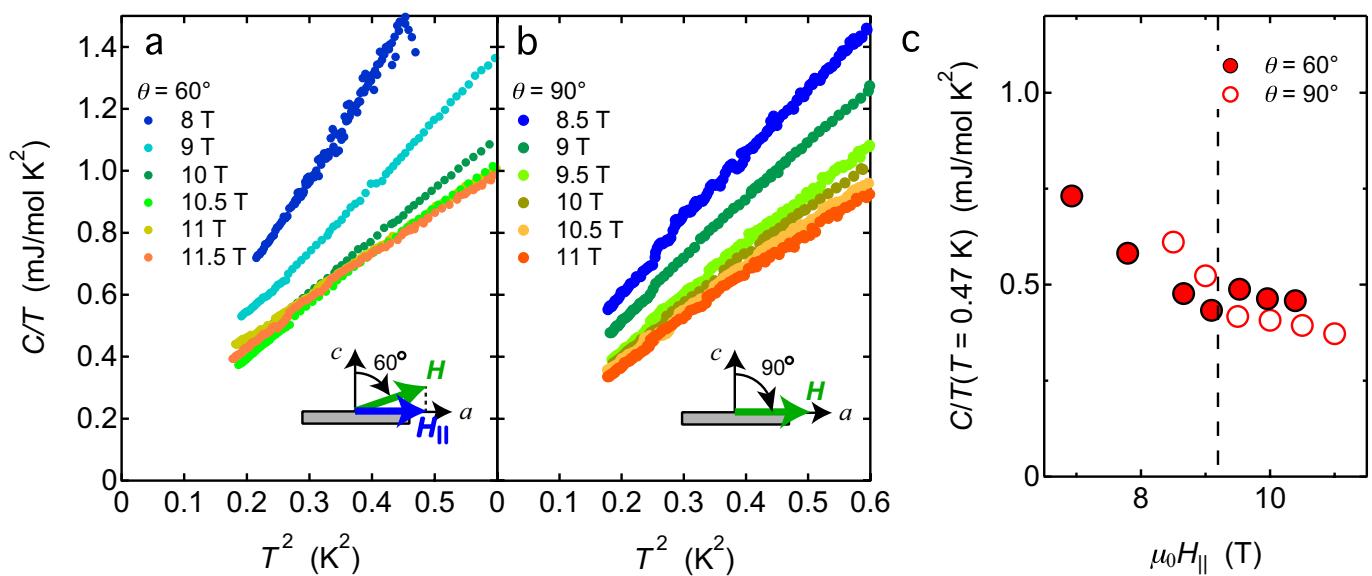
Extended Data Fig. 5 | Sample dependence of κ_{xy} . **a**, κ_{xy}/T measured in a different crystal (sample 2) for $\theta = 60^\circ$ (see inset of Fig. 2a) at 4.3 K, plotted as a function of H_\perp . The right scales represent the 2D thermal Hall conductance, κ_{xy}^{2D}/T , in units of $(\pi/6)(k_B^2/h)$. The half-integer thermal Hall conductance plateau is observed at $4.5 \text{ T} < \mu_0 H_\perp < 5.0 \text{ T}$. The field

where the overshoot behaviour from the quantization value is observed is slightly higher than that of sample 1, but the field where κ_{xy}/T vanishes ($\mu_0 H_\parallel \approx 9.3 \text{ T}$) is close to that of sample 1. **b**, κ_{xy}/T of sample 2 in a field tilted at $\theta = 60^\circ$, plotted as a function of H_\perp at 11 K. Error bars represent one standard deviation.



Extended Data Fig. 6 | Field dependence of thermal Hall conductivity in tilted fields at high temperatures. **a–d**, Thermal Hall conductivity, κ_{xy}/T , in a field tilted at $\theta = 60^\circ$ (**a**, **b**) and 45° (**c**, **d**), plotted as a function of H_\perp , (see inset of Fig. 2a). The right scales represent the 2D thermal Hall

conductance, κ_{xy}^{2D}/T , in units of $(\pi/6)(k_B^2/\hbar)$. Violet dashed lines represent the half-integer thermal Hall conductance, $\kappa_{xy}^{2D}/[T(\pi/6)(k_B^2/\hbar)] = 1/2$. Error bars represent one standard deviation.



Extended Data Fig. 7 | Specific heat above H_{\parallel}^* . **a, b**, Temperature dependence of C/T for $\theta = 60^\circ$ (**a**; H is tilted within the $a-c$ plane) and 90° (**b**). **c**, C/T at 0.47 K plotted as a function of H_{\parallel} for $\theta = 60^\circ$ and 90° . $C(H)/T$

exhibits a dip-like anomaly for $\theta = 60^\circ$ and a kink for $\theta = 90^\circ$ at $\mu_0 H_{\parallel} \approx 9.2 \text{ T}$ (dashed line). This field almost coincides with the characteristic field at which κ_{xy}/T vanishes.

Metallic nanoparticle contacts for high-yield, ambient-stable molecular-monolayer devices

Gabriel Puebla-Hellmann^{1,2,*}, Koushik Venkatesan^{3,4}, Marcel Mayor^{2,5,6} & Emanuel Lörtscher¹

Accessing the intrinsic functionality of molecules for electronic applications^{1–3}, light emission⁴ or sensing⁵ requires reliable electrical contacts to those molecules. A self-assembled monolayer (SAM) sandwich architecture⁶ is advantageous for technological applications, but requires a non-destructive, top-contact fabrication method. Various approaches ranging from direct metal evaporation⁶ over poly(3,4-ethylenedioxythiophene) polystyrene sulfonate⁷ (PEDOT:PSS) or graphene⁸ interlayers to metal transfer printing⁹ have been proposed. Nevertheless, it has not yet been possible to fabricate SAM-based devices without compromising film integrity, intrinsic functionality or mass-fabrication compatibility. Here we develop a top-contact approach to SAM-based devices that simultaneously addresses all these issues, by exploiting the fact that a metallic nanoparticle can provide a reliable electrical contact to individual molecules¹⁰. Our fabrication route involves first the conformal and non-destructive deposition of a layer of metallic nanoparticles directly onto the SAM (itself laterally constrained within circular pores in a dielectric matrix, with diameters ranging from 60 nanometres to 70 micrometres), and then the reinforcement of this top contact by direct metal evaporation. This approach enables the fabrication of thousands of identical, ambient-stable metal–molecule–metal devices. Systematic variation of the composition of the SAM demonstrates that the intrinsic molecular properties are not affected by the nanoparticle layer and subsequent top metallization. Our concept is generic to densely packed layers of molecules equipped with two anchor groups, and provides a route to the large-scale integration of molecular compounds into solid-state devices that can be scaled down to the single-molecule level.

The use of intrinsic molecular functionality is an attractive concept for the generation of novel electronic, photonic or sensing devices that provide responses not available using the silicon-based technologies of today. In the case of single-molecule junctions, however, their behaviour is directly dependent on the exact configuration of the junction, and the subatomic accuracy required for reproducibility cannot be achieved with current fabrication techniques. By contrast, the metal–SAM–metal architecture constrains the junction configurations and creates an ensemble average, which potentially mitigates variation between devices. Furthermore, a layer-by-layer approach, in which self-assembly provides the subatomic resolution in the junction direction, enables such devices to be fabricated with conventional methods.

The crucial challenge in a SAM-based approach is the formation of the top contact. This step should maintain film integrity as well as intrinsic functionality, while being compatible with mass fabrication and scalable in terms of active area. Various strategies have been proposed, ranging from direct evaporation onto cooled films in nanopores⁶ over PEDOT:PSS⁷ and graphene⁸ contact layers, through to EGaIn liquid-metal top electrodes¹¹, micro-contact printing⁹ and metal-transfer methods¹². Each strategy has drawbacks—for example, low yield owing to metal-filament penetration of the SAM^{13,14} or high series resistance¹⁵—and as such none of them can concurrently fulfil all of the abovementioned requirements.

To address all of these issues, we use metallic nanoparticles bound from solution to the top anchor group of the SAM. Known to establish a reliable electrical contact¹⁰, these particles provide a metallic, conformal and protective layer. The device geometry, schematically shown in Fig. 1a, is based on a platinum bottom electrode, onto which the molecular monolayer is assembled (Fig. 2b). A dielectric matrix constrains the molecules laterally into circular pores of variable diameter, from 70 μm down to 60 nm. Nanoparticles bind to the top anchor group of the molecular layer from solution (Fig. 1c), creating a film of immobilized particles. This top contact is subsequently reinforced by direct metal evaporation (Fig. 1d), which seals the device.

Platinum is chosen as the bottom-contact material because it is compatible with complementary metal–oxide–semiconductor (CMOS) fabrication. This choice does not affect the quality of the SAMs¹⁶, and the contact resistance of sulfur anchors to platinum is lower than that to gold, which is often used as the contact material¹⁷. We use conventional physical vapour deposition to avoid uncommon fabrication methods. This results in a roughness of 0.4 nm RMS (root mean square) over 1 μm^2 , which can be further improved by a factor of 2 using wafer-scale template stripping¹⁸.

The prototypical densely packed SAM consists of alkanethiol molecules comprising an alkane backbone with n carbon units and one (monothiol) or two (dithiol) terminal groups. As alkanes rapidly form robust and well-ordered monolayers, they have been used in most SAM-based platforms and are useful for benchmarking. We use alkanedithiols here, as the nanoparticle layer needs a top anchor group to adhere to. The second anchor group, however, introduces additional phases, such as the lying-down or the looped phase¹⁹, and can result in multilayer formation. Appropriate assembly conditions can reduce these issues^{19–21}. We study backbones with 4 to 10 carbon units, rather than 8 to 16 as is used typically. These shorter backbones enable an accurate assessment of the top-contact series resistance and the protective quality of the nanoparticles, albeit at the cost of a lower film quality²².

Gold nanoparticles are deposited onto the SAM from solution to form an initial metallic top-contact layer in a non-destructive and conformal manner²³. A non-polar solvent is required to enable contact between the liquid interface and the SAM²⁴. Agglomeration of nanoparticles in the solvent is suppressed by weakly bound ligands, which can be displaced locally to enable the formation of a chemical bond between the nanoparticle and the top anchor group of the molecules. This S–Au bond immobilizes the particles and creates an electrical contact between the metal and the molecule. The use of spherical particles reduces the area of molecules in contact with the top electrode: for 3-nm particles it is reduced by a factor of approximately 100, and can be optimized further by using smaller particles that are cubic in shape.

By comparing atomic force microscopy scans (Fig. 1e) of an empty pore (dashed black curve) and of a pore with the SAM (approximately 2 nm thickness) and nanoparticles (approximately 3 nm diameter) assembled (solid blue curve), we find an average depth difference between the two scans of around 5 nm. The nanoparticles preferentially

¹IBM Research - Zurich, Rüschlikon, Switzerland. ²Department of Chemistry, University of Basel, Basel, Switzerland. ³Department of Molecular Sciences, Macquarie University, North Ryde, New South Wales, Australia. ⁴Department of Chemistry, University of Zurich, Zurich, Switzerland. ⁵Karlsruhe Institute of Technology, Institute of Nanotechnology, Karlsruhe, Germany. ⁶Lehn Institute of Functional Materials, School of Chemistry, Sun Yat-Sen University, Guangzhou, China. *e-mail: gpu@zurich.ibm.com

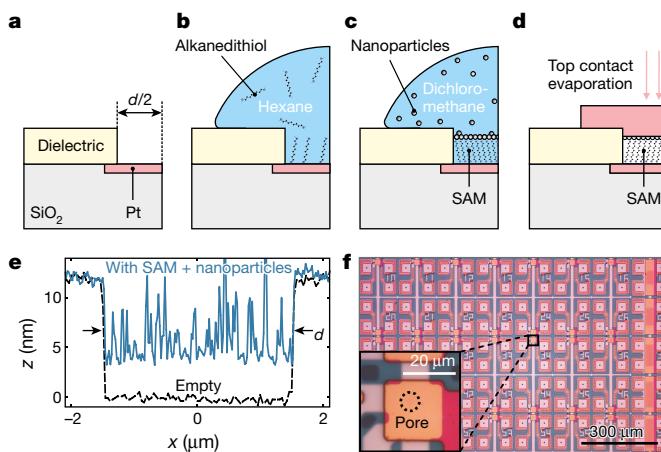


Fig. 1 | Device fabrication. **a**, Pores with a diameter d are etched into a dielectric film on a platinum electrode. **b**, Assembly of the SAM selectively on the electrode. **c**, Solution-based deposition of nanoparticles onto the SAM. **d**, Top contact evaporation. **e**, Atomic force microscopy scans of an empty pore (black, dashed) and of a pore in which SAM and nanoparticles are assembled (blue, solid). **f**, Optical image of device arrays. The inset shows a single device, with the location of a micrometre-sized pore indicated by the circle.

assemble on the SAM rather than the dielectric (see Methods), with a thickness that varies between one and two monolayers. A thin metal layer of approximately 20 nm is subsequently added by physical vapour deposition, with the nanoparticles probably acting as a physical shield and dissipating the kinetic energy of the metal atoms. This layer creates electrical contact both to and among the particles while hermetically sealing the pore (Fig. 1d). An optical image of part of a chip is shown in Fig. 1f, with the inset showing a single pore. In this work, gold nanoparticles are used because of the wide availability of mono-disperse solutions, but they can readily be replaced by platinum or palladium particles for CMOS-material compatibility. Here, fabrication is performed on 100-mm wafers up to the point of film assembly. In principle, however, the full process can be scaled up to larger wafer dimensions.

The individual molecular devices were characterized by their current–voltage (I – V) curves, obtained with triangle bias sweeps, starting at 0 V and ranging from 1.0 V to -1.0 V, respectively. Individual chips accommodate 62 arrays with 61 different pore diameters each, of which at least 54 arrays per chip are characterized. A maximum of 3,782 devices and a logarithmic scaling of the active area enable a statistical evaluation.

We initially discuss data from devices incorporating 1,10-decanedithiols. An I – V density plot combining the raw data of the three different pore diameters (790 nm, 1.8 μ m and 5.5 μ m) is shown in Fig. 2a (no filtering or averaging). The figure combines three density plots, with each colour (yellow, orange and blue) corresponding to the data obtained for one pore size (overlap is colored grey). We observe three distinct bundles with little overlap and a variation of less than half an order of magnitude, which demonstrates the reproducibility and low fluctuations of our approach. Figure 2b shows variable-temperature measurements for one device from each bundle. Upon cooling to 78 K, the current changes only slightly (less than 20%), which is consistent with measurements performed on similar metal–molecule–metal devices²⁵. We therefore conclude that non-resonant electron tunnelling is maintained as the underlying transport mechanism through the alkanedithiol monolayer²⁵.

Next, we focused on the relationship between the current I and the active area A . The density plot of the current at a bias of 0.2 V against the active area is shown in Fig. 2c. Each column represents data from 55 devices with nominally identical pore diameter. Histograms at diameters of 0.1, 1.0, 9.8 and 53 μ m are shown in Fig. 2d. Low scatter, negligible outliers, and a linear dependence between I and A are found, with

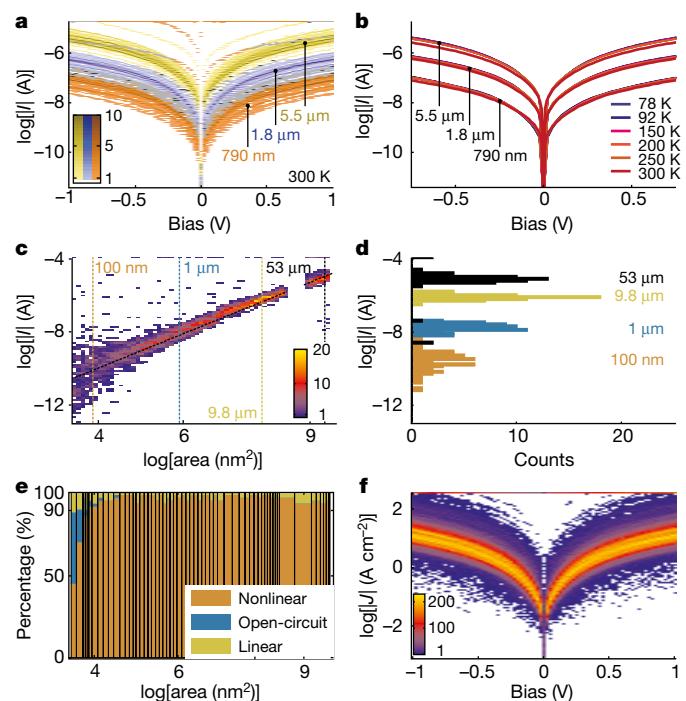


Fig. 2 | Transport properties of Pt-1,10-decanedithiol-Au junctions. **a**, I – V density plot for three pore diameters (790 nm, 1.8 μ m and 5.5 μ m), comprising 55 devices each. **b**, Temperature-dependent I – V curves of one individual device for each pore diameter. **c**, I – A density plot over all pore sizes with 55 devices per size, extracted at a bias of 0.2 V. **d**, Histograms of I for devices with four selected pore diameters. **e**, Percentage of devices categorized as nonlinear, open-circuit or linear as a function of active area, A . **f**, J – V density plot of all 3,204 nonlinear devices.

A ranging from 4.1×10^9 nm 2 down to 5×10^3 nm 2 (the gap visible around 10^9 nm 2 is due to an area range not covered in the design). We attribute the increased scatter for $A < 3 \times 10^5$ nm 2 to variations in the active area caused by imperfect etching of the dielectric.

Individual I – V curves can be categorized as open-circuit, linear or nonlinear (see Methods for the exact definition and examples), with open-circuit curves mainly representing pores that are not properly etched or those with broken electrodes, and linear curves representing those in which top and bottom electrode are shorted. As transport through an alkanedithiol molecular layer typically yields a nonlinear I – V curve, the nonlinear class contains the curves of the devices in which proper contact to the layer was achieved. Figure 2e shows a graph of this class distribution as a function of A . An open-circuit fraction exceeding 10% is observed for only the two smallest pore diameters, which again indicates imperfect etching. For $A > 10^5$ nm 2 , however, curves are classified only as linear or nonlinear. Over the entire range of A , a total of 3,204 devices (95.5%) are classified as nonlinear, 67 as open circuit (2%) and 84 as linear (2.5%), which proves that the nanoparticle layer successfully inhibits the formation of SAM-penetrating metal filaments even for large values of A , probably by physically shielding the SAM.

In addition to a high device yield, reproducible device-to-device transport properties are also mandatory for an application-oriented molecular platform. Figure 2f shows a density plot of the calculated current density, J , against voltage, V , for the 3,204 devices classified as nonlinear. A narrow distribution with a value of 11.1 ± 7.1 A cm $^{-2}$ at 1.0 V is observed, the result of both ensemble averaging and a reduction in possible contact geometries, with the absolute value influenced by the nanoparticle diameter (see Methods). Furthermore, the devices are stable up to at least ± 1.0 V in bias, do not require any conditioning, and retain their transport properties for several months even when stored under ambient conditions (see Methods). A comprehensive comparison of the device characteristics achieved in other platforms is provided in Methods. Despite the well-defined current, it is difficult

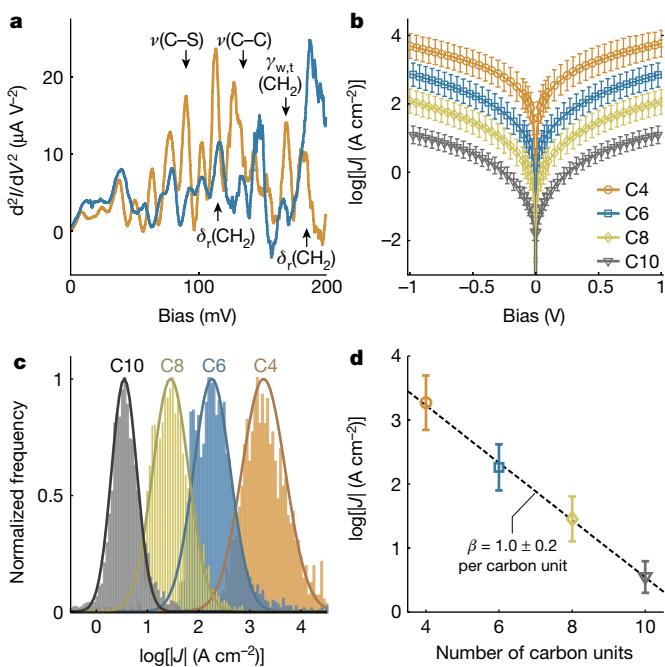


Fig. 3 | Molecular origin of the transport properties. **a**, Vibrational spectra obtained from IETS of 1,10-decanedithiol devices with pore diameters of 7.6 μm (blue trace) and 8.5 μm (orange trace) at 4 K, with selected molecular vibrational peaks labelled. **b**, Mean current density, J , for C4, C6, C8 and C10 alkanedithiols. **c**, Histograms of J extracted at a bias of 0.5 V and corresponding Gaussian fits. **d**, Average J at a bias of 0.5 V against backbone chain length and corresponding fit of the tunnelling coefficient β .

to extrapolate an average current per molecule owing to the unknown molecular and nanoparticle packing densities.

The data presented demonstrate nonlinear responses with high yield and high device-to-device reproducibility. However, in principle, these characteristics could have an alternative origin that is distinct from transport through the SAM. To confirm the molecular origin of the device response, we performed inelastic electron tunnelling spectroscopy (IETS) and studied the dependence of transport on the alkane length. Figure 3a shows spectra of two devices, with pore diameters of 7.6 μm and 8.5 μm , acquired at 4.2 K. Various vibrational peaks are present, consistent with previous studies^{26,27}. Several peaks present in both traces correspond to characteristic molecular vibrations, namely $\nu(\text{C-S})$ (C-S stretch; 92 meV), $\delta_r(\text{CH}_2)$ (CH₂ rock; 115 meV), $\nu(\text{C-C})$ (C-C stretch; 135 meV), $\gamma_{\text{w,t}}\text{CH}_2$ (CH₂ wag; 160 meV) and $\delta_s(\text{CH}_2)$ (CH₂ scissor; 185 meV). This inelastic electron-scattering signal provides evidence that transport is dominated by the SAM, although further studies will be necessary to identify all of the peaks observed.

Additional evidence for the molecular origin of the transport properties was provided by comparing devices with different alkane lengths (4, 6, 8 and 10 carbons, labelled C4, C6, C8 and C10, respectively); this also demonstrates the flexibility of the platform in terms of compound dimensions. Figure 3b compares the mean and standard deviation of J , obtained for each compound by fitting a Gaussian to the histogram at each bias point. The curves are similar in functional behaviour, with a shift in J and a decreased deviation for C10. An example histogram of J extracted at 0.5 V bias (Fig. 3c) shows four distinct and Gaussian-like distributions. In a simple model, the SAM is considered a tunnel barrier with $J \propto 10^{\beta d/2}$ ^{30,31}, where β is the decay coefficient and d is the barrier width²⁸. Extracting β from our data yields $\beta_n = 1.0 \pm 0.2$ per carbon (Fig. 3d), in agreement with literature values for through-bond tunnelling^{13,22,29}. This length dependence shows not only that the molecular layer maintains its physical and chemical integrity, but also that the SAM dominates the transport characteristics and is not masked by the contact resistance or influenced by ligand-mediated series effects, and that Coulomb blockade effects³⁰ can be ruled out.

The molecular integration approach presented offers a simple and non-destructive top-contact formation step, based on nanoparticles chemisorbed to the top anchor groups of densely packed molecules. Our dielectric micro- and nanopore platform further provides control over the active area and is flexible in terms of molecular length. This flexibility, together with automated characterization of thousands of devices, enables rapid screening of molecular compounds in two-terminal junctions. Our approach will enable the accelerated development of chemically tunable electronic building blocks, such as conductance-switching memories, artificial synapses and quantum-interference devices.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0275-z>.

Received: 21 March 2017; Accepted: 4 May 2018;

Published online 11 July 2018.

- Elbing, M. et al. A single-molecule diode. *Proc. Natl Acad. Sci. USA* **102**, 8815–8820 (2005).
- Chen, J., Reed, M. A., Rawlett, A. M. & Tour, J. M. Large on-off ratios and negative differential resistance in a molecular electronic device. *Science* **286**, 1550–1552 (1999).
- Schwarz, F. et al. Field-induced conductance switching by charge-state alternation in organometallic single-molecule junctions. *Nat. Nanotechnol.* **11**, 170–176 (2016).
- Marquardt, C. W. et al. Electroluminescence from a single nanotube–molecule–nanotube junction. *Nat. Nanotechnol.* **5**, 863–867 (2010).
- Ponce, J. et al. Effect of metal complexation on the conductance of single-molecular wires measured at room temperature. *J. Am. Chem. Soc.* **136**, 8314–8322 (2014).
- Zhou, C., Deshpande, M. R., Reed, M. A., Jones, L. & Tour, J. M. Nanoscale metal/ self-assembled monolayer/metal heterostructures. *Appl. Phys. Lett.* **71**, 611–613 (1997).
- Akkerman, H. B., Blom, P. W. M., de Leeuw, D. M. & de Boer, B. Towards molecular electronics with large-area molecular junctions. *Nature* **441**, 69–72 (2006).
- Wang, G., Kim, Y., Choe, M., Kim, T.-W. & Lee, T. A new approach for molecular electronic junctions with a multilayer graphene electrode. *Adv. Mater.* **23**, 755–760 (2011).
- Loo, Y.-L., Lang, D. V., Rogers, J. A. & Hsu, J. W. P. Electrical contacts to molecular layers by nanotransfer printing. *Nano Lett.* **3**, 913–917 (2003).
- Cui, X. D. et al. Reproducible measurement of single-molecule conductivity. *Science* **294**, 571–574 (2001).
- Nijhuis, C. A., Reus, W. F., Barber, J. R., Dickey, M. D. & Whitesides, G. M. Charge transport and rectification in arrays of SAM-based tunneling junctions. *Nano Lett.* **10**, 3611–3619 (2010).
- Jeong, H. et al. A new approach for high-yield metal–molecule–metal junctions by direct metal transfer method. *Nanotechnology* **26**, 025601 (2015).
- Lee, T. et al. Comparison of electronic transport characterization methods for alkanethiol self-assembled monolayers. *J. Phys. Chem. B* **108**, 8742–8750 (2004).
- Kim, T.-W., Wang, G., Lee, H. & Lee, T. Statistical analysis of electronic properties of alkanethiols in metal–molecule–metal junctions. *Nanotechnology* **18**, 315204 (2007).
- Neuhausen, A. B., Hosseini, A., Sulpizio, J. A., Chidsey, C. E. D. & Goldhaber-Gordon, D. Molecular junctions of self-assembled monolayers with conducting polymer contacts. *ACS Nano* **6**, 9920–9931 (2012).
- Lee, S. et al. Self-assembled monolayers on Pt(111): molecular packing structure and strain effects observed by scanning tunnelling microscopy. *J. Am. Chem. Soc.* **128**, 5745–5750 (2006).
- Beebe, J. M., Engelkes, V. B., Miller, L. L. & Frisbie, C. D. Contact resistance in metal–molecule–metal junctions based on aliphatic SAMs: effects of surface linker and metal work function. *J. Am. Chem. Soc.* **124**, 11268–11269 (2002).
- Puebla-Hellmann, G., Mayor, M. & Lörtscher, E. Ultraflat nanopores for wafer-scale molecular-electronic applications. In *2015 IEEE 15th International Conference on Nanotechnology (IEEE-NANO)* 1197–1201 (IEEE, 2015).
- Akkerman, H. B. et al. Self-assembled-monolayer formation of long alkanedithiols in molecular junctions. *Small* **4**, 100–104 (2008).
- Jia, J. et al. Lying-down to standing-up transitions in self assembly of butanedithiol monolayers on gold and substitutional assembly by octanethiols. *J. Phys. Chem. C* **117**, 4625–4631 (2013).
- Chah, S., Fendler, J. H. & Yi, J. *In-situ* analysis of stepwise self-assembled 1,6-hexanedithiol multilayers by surface plasmon resonance measurements. *Chem. Commun.* **18**, 2094–2095 (2002).
- Engelkes, V. B., Beebe, J. M. & Frisbie, C. D. Length-dependent transport in molecular junctions based on SAMs of alkanethiols and alkanedithiols: effect of metal work function and applied bias on tunnelling efficiency and contact resistance. *J. Am. Chem. Soc.* **126**, 14287–14296 (2004).

23. Lörtscher, E., Mayor, M. & Puebla-Hellmann, G. Contacting molecular components. US patent application 20180062076 (2018).
24. Sur, U. K. & Lakshminarayanan, V. Existence of a hydrophobic gap at the alkanethiol SAM-water interface: an interfacial capacitance study. *J. Colloid Interface Sci.* **254**, 410–413 (2002).
25. Wang, W., Lee, T. & Reed, M. A. Mechanism of electron conduction in self-assembled alkanethiol monolayer devices. *Phys. Rev. B* **68**, 035416 (2003).
26. Wang, W., Lee, T., Kretzschmar, I. & Reed, M. A. Inelastic electron tunneling spectroscopy of an alkanedithiol self-assembled monolayer. *Nano Lett.* **4**, 643–646 (2004).
27. Jeong, H. et al. Investigation of inelastic electron tunneling spectra of metal-molecule-metal junctions fabricated using direct metal transfer method. *Appl. Phys. Lett.* **106**, 063110 (2015).
28. Simmons, J. G. Low-voltage current-voltage relationship of tunnel junctions. *J. Appl. Phys.* **34**, 238–239 (1963).
29. Simeone, F. C. et al. Defining the value of injection current and effective electrical contact area for EGaIn-based molecular tunneling junctions. *J. Am. Chem. Soc.* **135**, 18131–18144 (2013).
30. Morita, T. & Lindsay, S. Determination of single molecule conductances of alkanedithiols by conducting-atomic force microscopy with large gold nanoparticles. *J. Am. Chem. Soc.* **129**, 7262–7263 (2007).

Acknowledgements We acknowledge technical support from M. Tschudy, S. Reidt, M. Sousa, U. Drechsler, A. Zulji and M. Bürge, as well as strategic support from B. Michel, W. Riess and A. Curioni. This work was funded by the NCCR MSE and SNF NRP 62.

Author contributions G.P.-H. conceived, developed and performed device fabrication, performed data collection and analysis, and wrote the manuscript. K.V. and M.M. provided chemical support and commented on the manuscript. E.L. initiated and supervised the project, contributed to the experimental setups and data analysis and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0275-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Device fabrication. Device fabrication was based on standard 4" silicon wafers with a 150–200-nm thick thermal oxide layer. Bottom electrodes consisting of 60 nm Pt with a 0.5-nm Ti adhesion layer were defined on the wafer by optical photolithography, electron-beam evaporation (BAK501LL, Evatec) and lift-off. To deposit the dielectric, a two-step approach was used: first, an adhesion layer of titanium oxide (4 cycles) and a layer of silicon nitride (50 cycles, around 3 nm) were deposited by atomic layer deposition (FlexAL, Oxford Instruments), yielding a low-quality protective layer. Second, plasma enhanced chemical vapour deposition (PECVD; PlasmaPro 100 PECVD, Oxford Instruments) was used to deposit 10 nm of high-quality silicon nitride.

Pores were etched in the dielectric using two methods: first, large features were defined by photolithography and etched using buffered hydrofluoric acid (BHF). Second, micron and submicron features were patterned by electron-beam lithography and a combination of reactive-ion etching (CHF₃/O₂, NGP 80, Oxford Instruments) and wet etching (BHF). The wafers were then cleaned in piranha etch (3:1 H₂SO₄:H₂O₂) and cleaved using a silicon scribe. Individual chips were cleaned with plasma (120 s Ar/H plasma, 900 s O plasma, TePla 100-E) before being immersed in solutions of the different molecular species. Dedicated glassware was used for the molecular solutions and plasma-cleaned before use. Molecular solutions were 50 mM (C4/C6: 20 μl, C8/C10: 30 μl in 3 ml hexane), freshly prepared using commercially available alkanedithiols (Sigma Aldrich (C4,C6,C8), Alfa Aesar (C10)) and hexane (Sigma Aldrich) stored under argon. The glassware was then back-filled with argon, sealed and stored in a light-shielded box for 2–3 h (C4, C6) or 20–24 h (C8, C10). After rinsing twice with hexane and drying with nitrogen, the chips were placed in a 0.1 g l⁻¹ solution of nanoparticles (3-nm, PVP, NanoPartz) in dichloromethane (Sigma Aldrich) for 90 s and then dried with nitrogen.

The 20-nm-thick gold top electrode was deposited by electron-beam evaporation, with a rate ramping from 0 to 0.2 nm s⁻¹ over approximately 20 s. The electrode was then patterned by photolithography using a standard resist (MicroChemicals AZ6612), omitting any baking steps, and subject to ion milling (Ionfab 300, Oxford Instruments). Chips were then cleaned with oxygen plasma (PVA TePla Gigabatch, 200 W, 120 s) and rinsed in acetone/isopropanol to remove the resist.

Data acquisition and processing. Devices were characterized at room temperature and ambient conditions using an automated probe station (Cascade Summit 12000) combined with a semiconductor parameter analyser (HP B1500A). Temperature-dependent measurements were performed in vacuum in probe station cooled with liquid nitrogen (Janis). The temperature was controlled using a controller (Lakeshore) and a heating element. For IETS measurements, samples were bonded to a chip carrier, mounted in a home-made dip stick and immersed in liquid helium. The DC and AC (337.3 Hz, 4 mV RMS) components of the signal, generated by a lock-in amplifier (Stanford Instruments, SR830), were combined and low-pass (1 kHz) filtered through an adder/filter (Stanford Instruments, SR560) and applied to the sample. The second harmonic signal was detected by the same lock-in amplifier.

All *I*–*V* data are presented as measured without any post-processing. The current density *J* was calculated using the active area as determined by scanning electron microscopy. IETS measurements were smoothed using five-point boxcar averaging.

I–V characterization. For datasets with several thousand *I*–*V* curves, an automated routine was required to differentiate between different types of device behaviour. We differentiated between three types: open circuit, linear and nonlinear. The categorization algorithm works as follows: first, all curves with an average absolute current below 25 pA are categorized as open circuit; three examples are shown in Extended Data Fig. 1a. The cut-off value is above our noise floor (around 10 pA) and also removes samples with low signal-to-noise (solid blue curve), where differentiation between linear and nonlinear is difficult. For the remaining curves, the algorithm calculates the deviation, δ , between the resistances at 50 mV and 300 mV: $\delta = |1 - R_{50 \text{ mV}}/R_{300 \text{ mV}}|$. Curves with $\delta < 0.1$ are categorized as linear, the others as nonlinear. Sample curves for different δ are shown in Extended Data Fig. 1b, in which the top two are categorized as linear and the bottom two as nonlinear. Note that the current is normalized for a better comparison. The two bias values are chosen such that the lower point still yields good signal-to-noise for small pores and the upper value such that shorted curves are not in current compliance, which would lead to a false categorization as nonlinear. The cut-off value of 0.1 is chosen such that a smaller nonlinearity, for example due to Joule heating of a short, is still categorized as linear.

Characterization of the bottom electrode. The surface morphology of the electrode is an important factor influencing the assembly of the SAM and therefore the properties of the entire device. In particular, deposition and etching of the dielectric film on top of the electrode should not degrade the electrode surface. The surface roughness was investigated using atomic force microscopy (Veeco Dim V, tapping mode with Nanosensors Pointprobe+ tips). A pore defined by electron beam

lithography and reactive ion etching is shown in Extended Data Fig. 2a, the small particles visible on the dielectric areas around the pore are likely to be remaining resist. A smaller scan of the exposed electrode surface is shown in Extended Data Fig. 2b, from which a roughness of 0.39 nm RMS over 1 μm² is extracted, similar to the roughness of the electrode before deposition and etching.

Nanoparticle film characterization. After deposition of the molecular monolayer and the nanoparticles, a similar pore was scanned, and is shown in Extended Data Fig. 2c. The inside of the pore was considerably roughened and the pore had reduced depth. A smaller scan of the surface inside the pore is shown in Extended Data Fig. 2d and showed round features that were not present before, as expected for a nanoparticle layer. We observed an increase in roughness from 0.39 nm RMS to 3.18 nm RMS over 1 μm². As the lateral resolution is limited by convolution of the topology with the tip radius (7–10 nm radius), the film surface is difficult to image. We observed round structures of different dimensions, up to 10–15 nm height and 40–50 nm apparent diameter. Although the dispersion of the used nanoparticles is stated to be 20%, these protrusions may be gold, ligand agglomerates or both. In both cases, such features reduce the effective area and an optimization of the nanoparticle dispersion and deposition process will be required.

High-resolution transmission electron microscopy of a finished device. To investigate the interface between the nanoparticle film and the top electrode, transmission electron microscopy (TEM) was performed on a cross-section of a measured C10 device. A TEM lamella was extracted from the device using a focused gallium ion-beam system (FEI). After extraction, the lamella was thinned down to a width of approximately 40 nm and loaded into a TEM system (JEOL), in which images were taken at an acceleration voltage of 200 kV. An overview image is shown in Extended Data Fig. 3a, in which the SAM separates the top and bottom electrode (bright layers) on the left side. The slightly thicker separation on the right side is the SiN_x dielectric. The top electrode is mainly flat but shows warped regions, one of which is shown at higher magnification in Extended Data Fig. 3b. Here, the SAM region separating top and bottom electrode is clearly visible. We also observed small voids in the top electrode, which may have origins in the larger protrusions visible in the atomic force microscopy (AFM) scans. More importantly, we observed no apparent features reminiscent of nanoparticles, which suggests that the particles have fused with the top electrode.

Increasing the magnification yields close to atomic resolution, as can be seen in Extended Data Fig. 3c, d, in which the crystal lattices and grain boundaries of both electrodes become visible. However, no features on the scale of 3 nm were visible, which suggests that the ligands have been displaced and the particles have fused to create a uniform, albeit rough, top electrode. Note that owing to the low scattering coefficient of carbon, the alkanedithiol SAM is not visible in these images.

To ensure that the top electrode close to the SAM is indeed gold, we performed energy dispersive X-ray spectroscopy on the top electrode, in close (around 2 nm) vicinity to the SAM. The resulting spectrum is shown in Extended Data Fig. 4. Peaks were automatically identified by the software. Three elements were detected: carbon, gold and copper. The carbon and copper peaks are always present owing to background carbon in the chamber and copper from the sample holder. We therefore conclude that the top electrode consists mainly of gold and carbon and the particles visible in the AFM images are indeed either gold or carbon.

Temperature dependence. Transport through molecules can be the result of several different mechanisms, such as resonant or off-resonant tunnelling, hopping or thermionic emission²⁵. The temperature dependence of the transport properties can help to distinguish between these mechanisms. In the case of alkanedithiols, the conduction mechanism is expected to be off-resonant tunnelling³¹; for low bias, this can be described by the Simmons model and is expected to be temperature independent²⁸. As alkanethiol SAMs are often implemented in large-area molecular junctions, literature is available on the temperature-dependent transport properties of these devices. For metal–molecule–metal junctions, no notable temperature dependence was found for three different manufacturing approaches: nanopores²⁵, direct evaporation on larger pores¹⁴ and direct metal transfer¹². Similar results were obtained for PEDOT devices⁷, with the notable exception of one study, which attributed the dependence to the different polymer used¹⁵.

To demonstrate the weak temperature dependence in our devices, Extended Data Fig. 5a shows Arrhenius plots for two of the three devices shown in Fig. 1b (the third is omitted for clarity). Both devices show a slight temperature dependence with a tendency to higher current at lower temperature for higher bias, similar to a previously observed trend¹², and consistent in magnitude with other work^{7,14,25}.

Preliminary diameter dependence. The absolute value of the current density is probably determined by the molecular film–nanoparticle interface in our devices. Whereas typically 1–4 molecules will attach to a nanoparticle¹⁰, a larger number of molecules will not establish contact to the round particle. The fraction of uncontacted molecules will depend on the packing density and the size of the nanoparticles. The results of a preliminary study of the nanoparticle diameter are shown in Extended Data Fig. 5b, c. For both particle sizes, we observe a reduction in

current density for longer molecules, consistent with transport through the SAM. Furthermore, we observe a shift to lower current densities when using particles of 5 nm instead of 3 nm diameter, consistent with an expected 'shadowing effect', as larger particles will leave a larger fraction of molecules without contact.

Device repeatability and stability. In addition to the comparison between devices, single device stability and repeatability are also desirable. One aspect is the stability of the I - V curves obtained, as SAM-based molecular electronic devices often require several voltage sweeps for their properties to stabilize. Because each device is characterized by a triangle sweep, we can compare the upward and downward sweeps to characterize the single-sweep repeatability. A histogram of the current ratio between upward and downward sweeps at a bias of 0.5 V obtained for C10 devices is shown in Extended Data Fig. 6a. Out of a total of 3,355 devices, 3,171 are classified as nonlinear on both sweeps, compared with 3,204 on the upward sweep. The data presented incorporate only the devices classified as nonlinear on both sweeps. As visible from the graph, 58% (1,867/3,171) of the devices exhibit a current difference of less than 0.5%. Furthermore, 93% (2,949/3,171) are within 10%, which demonstrates excellent sweep stability.

A second aspect is long-term stability, for which a part of the C10 chip was measured a second time after storage under ambient conditions for 136 days. In contrast to PEDOT devices, which were initially characterized in vacuum and, after ambient storage, recover their initial properties only in vacuum³², the data presented was obtained under ambient conditions without any vacuum treatment. The ratio of the initial current to the current measured after 136 days is shown as a histogram in Extended Data Fig. 6b. A distinct peak is visible at a ratio of 1, indicating stable devices, and a second peak is visible close to 0.75. Half of the devices are in the interval between 0.6 and 1.4, even though the devices have been cycled several times between 77 K and 300 K. The second peak mainly originates from pores with an area exceeding 1,000 μm^2 and, although the reason for this consistent behaviour is currently unclear, it may relate to the very large active area and therefore increased number of defects in the SAM. As the top electrode in the current devices is comparatively thin (20 nm), we expect better stability in future devices with a thicker top layer.

Comparison with existing platforms. A large variety of different approaches have been attempted for the implementation of SAM-based devices. Extended Data Table 1 enables a comparison of the major approaches with the work presented regarding several aspects, such as the current density J , the standard deviation, the number of devices measured, the overall yield, the pore diameter and the mass-fabrication compatibility.

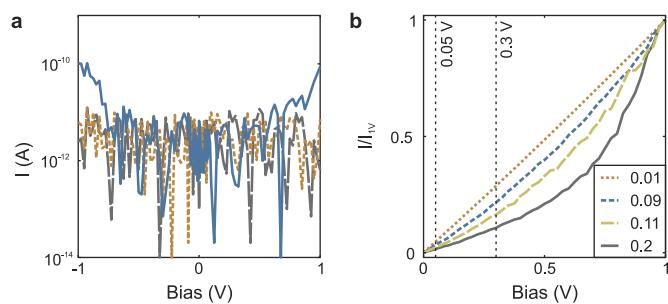
Our approach enables almost an order of magnitude more devices to be characterized over a considerably larger range of pore diameters at yields comparable to the best results obtained so far, all while maintaining mass-fabrication compatibility. We do observe a lower current density J compared to direct metal evaporation^{33,34}, due to the use of spherical nanoparticles which do not contact all the molecules in the SAM. However, we do obtain larger current densities than

PEDOT^{7,15,35}, reduced graphene oxide³⁶ and eutectic GaIn³⁷ and comparable densities to transfer printing^{9,38}. This parameter may be optimized further by the use of smaller and/or differently shaped particles.

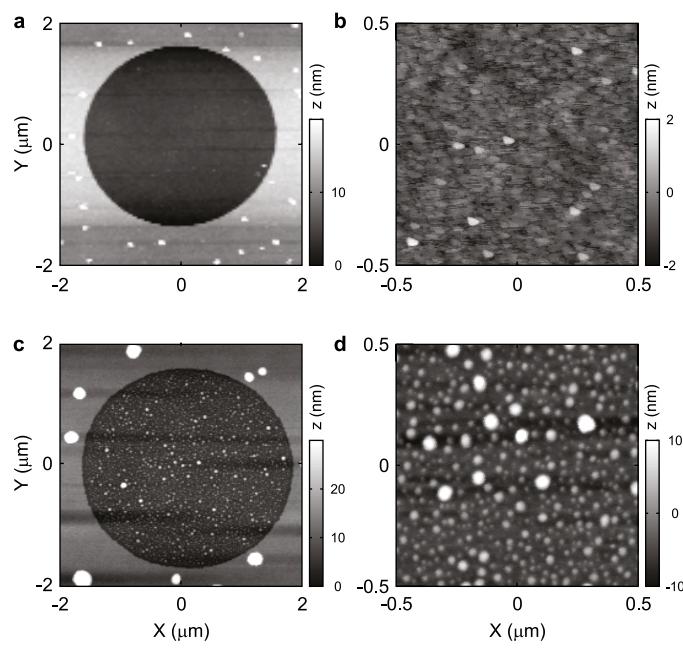
A key parameter is the deviation of the current density σ_J , which is plotted as a function of device area in Extended Data Fig. 7. Although we obtain an overall FWHM \log_{10} of 0.9, the plot shows that the scatter is considerably reduced for pore areas in the 10^2 – $10^3 \mu\text{m}^2$ range, a typical range for SAM-based devices. The obtained values in this region are comparable to the best values obtained in PEDOT devices. Furthermore, the plot shows the large range of device areas accessible with our method.

Additional information for devices based on C4, C6 and C8. In general, we observe similar behaviour for all four species of molecules investigated. However, the quality of the film decreases for decreasing chain length, which leads to an increased number of defects, larger scatter and lower stability of the final devices. Histograms of the current I versus active area A at a bias of 0.2 V are shown in Extended Data Fig. 8a–d as well as of the current density J versus bias (Extended Data Fig. 8i–l). We obtain excellent results for C6, C8 and C10, with typically 80% or more showing nonlinear behaviour, as shown in Extended Data Fig. 8e–h. For the shortest chain length, two slopes are visible in the I - A histogram, as linear curves are not excluded from the data shown, which also broadens the peaks in the J - V histogram. Although the yield is considerably reduced (Extended Data Fig. 8i), such short molecules are often not investigated at all. Here, we demonstrate that the approach presented can be also be used for films of typically low quality. **Data availability.** The data presented in this paper are available from the corresponding author upon reasonable request.

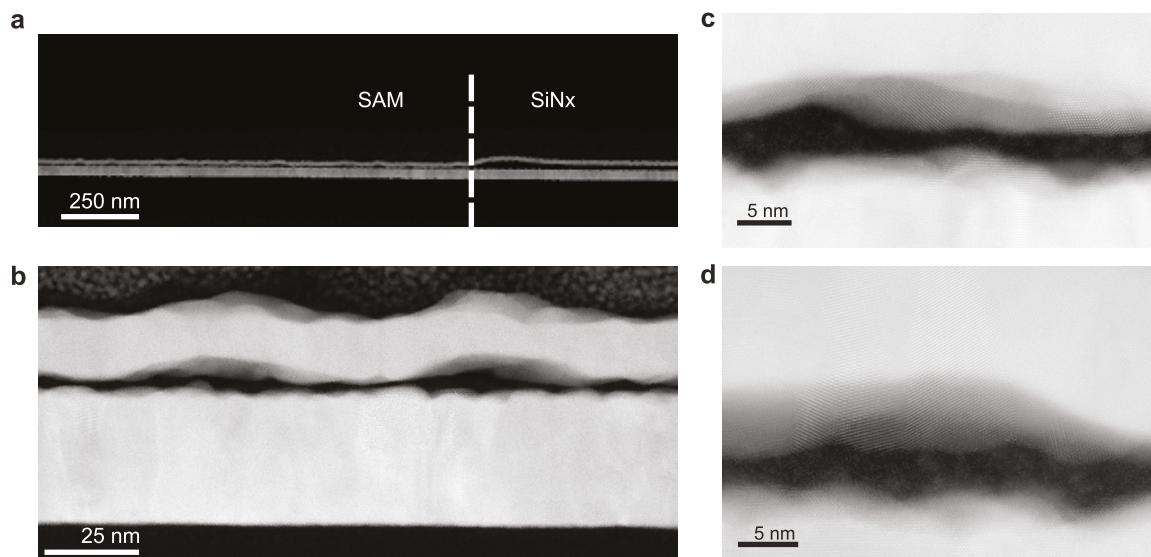
31. Wold, D. J., Haag, R., Rampi, M. A. & Frisbie, C. D. Distance dependence of electron tunneling through self-assembled monolayers measured by conducting probe atomic force microscopy: Unsaturated versus saturated molecular junctions. *J. Phys. Chem. B* **106**, 2813–2816 (2002).
32. Akkerman, H. B. *Large-area Molecular Junctions*. PhD thesis, Univ. Groningen (2008).
33. Wang, G., Kim, T.-W., Lee, H. & Lee, T. Influence of metal-molecule contacts on decay coefficients and specific contact resistances in molecular junctions. *Phys. Rev. B* **76**, 205320 (2007).
34. Lee, T., Wang, W. & Reed, M. A. Intrinsic electronic transport through alkanedithiol self-assembled monolayer. *Jpn. J. Appl. Phys.* **44**, 523 (2005).
35. Park, S. et al. Flexible molecular-scale electronic devices. *Nat. Nanotechnol.* **7**, 438–442 (2012).
36. Seo, S. et al. Solution-processed reduced graphene oxide films as electronic contacts for molecular monolayer junctions. *Angew. Chem. Int. Ed.* **51**, 108–112 (2012).
37. Reus, W. F. et al. Statistical tools for analyzing measurements of charge transport. *J. Phys. Chem. C* **116**, 6714–6733 (2012).
38. Jeong, H. et al. Statistical investigation of the length-dependent deviations in the electrical characteristics of molecular electronic junctions fabricated using the direct metal transfer method. *J. Phys. Condens. Matter* **28**, 094003 (2016).



Extended Data Fig. 1 | Device response classification. **a**, Three example curves of an open circuit. **b**, Four example curves with different ratios of resistance at 50 mV and 300 mV bias. The top two (orange and blue lines) are categorized as linear, whereas the lower two (yellow and grey lines) are categorized as nonlinear.

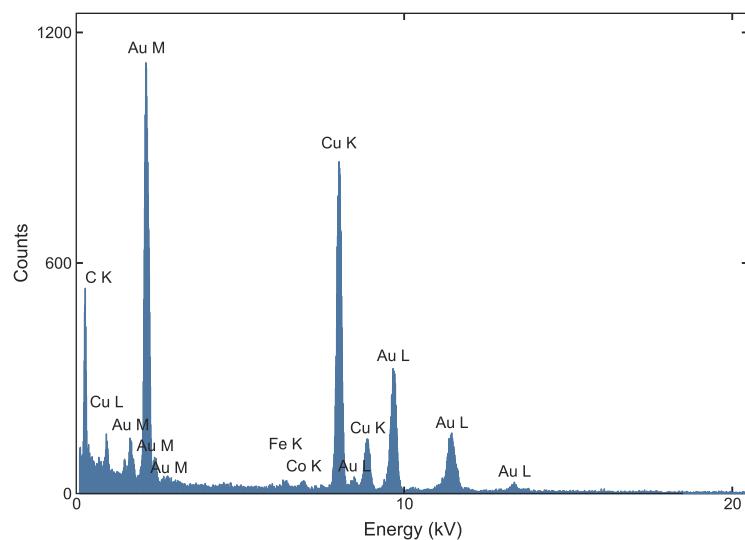


Extended Data Fig. 2 | Surface and nanoparticle film topology. **a**, AFM scan of a 2.7- μm -diameter pore. **b**, The surface at the bottom of the pore with a roughness of 0.39 nm RMS. **c**, A similar pore after film growth and nanoparticle deposition. **d**, Scan of the surface inside the pore, showing circular particles.

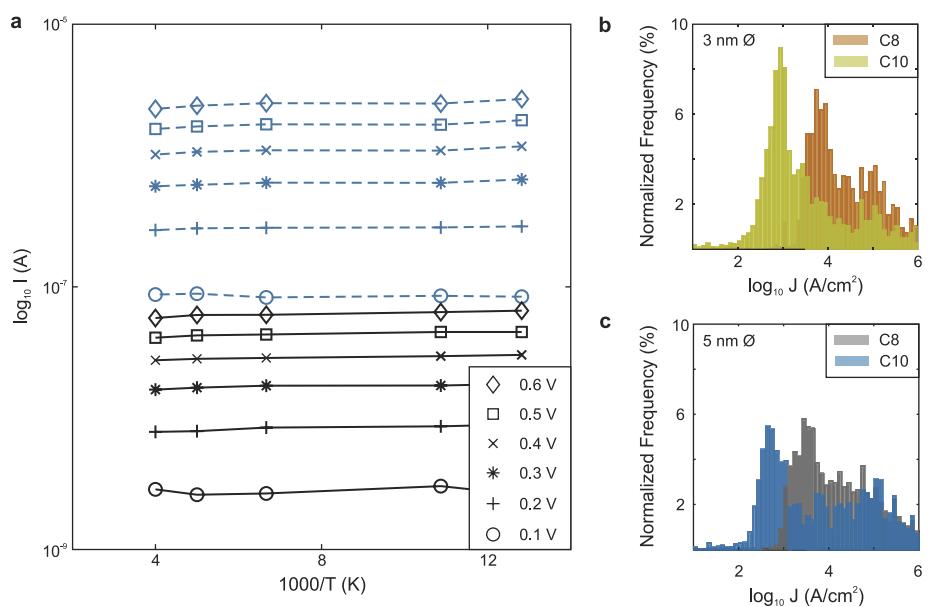


Extended Data Fig. 3 | Device cross sections. **a**, TEM images of part of a C10 device lamella. Both the platinum bottom electrode and the gold top electrode are visible as bright layers, separated by the SAM on the left side and the SiN_x dielectric on the right. The top electrode is warped in some regions. **b**, Zoom into a warped region: the bottom electrode, SAM region

and top electrode are clearly visible, as well as the electron-deposited platinum protection layer for extraction of the lamella (the grainy area at the top of the image). **c, d**, High-resolution TEM images of two different areas of the SAM region. Although crystal lattices are visible, no 3-nm-sized features can be identified that would correspond to nanoparticles.

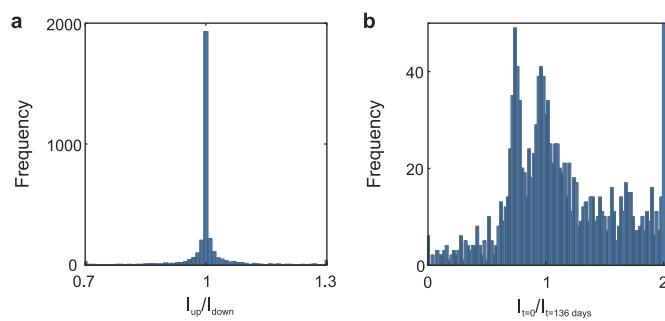


Extended Data Fig. 4 | Element analysis. Energy-dispersive X-ray spectrum of a region located approximately 2 nm above the SAM. Three elements can be identified: carbon, gold and copper.

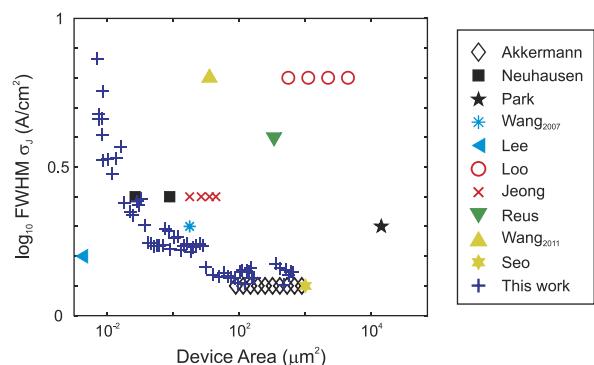


Extended Data Fig. 5 | Dependence of device characteristics on temperature and particle diameter. **a**, Arrhenius plot for two devices (solid black, 790 nm diameter; dashed blue, 5.5 μm), showing a very slight

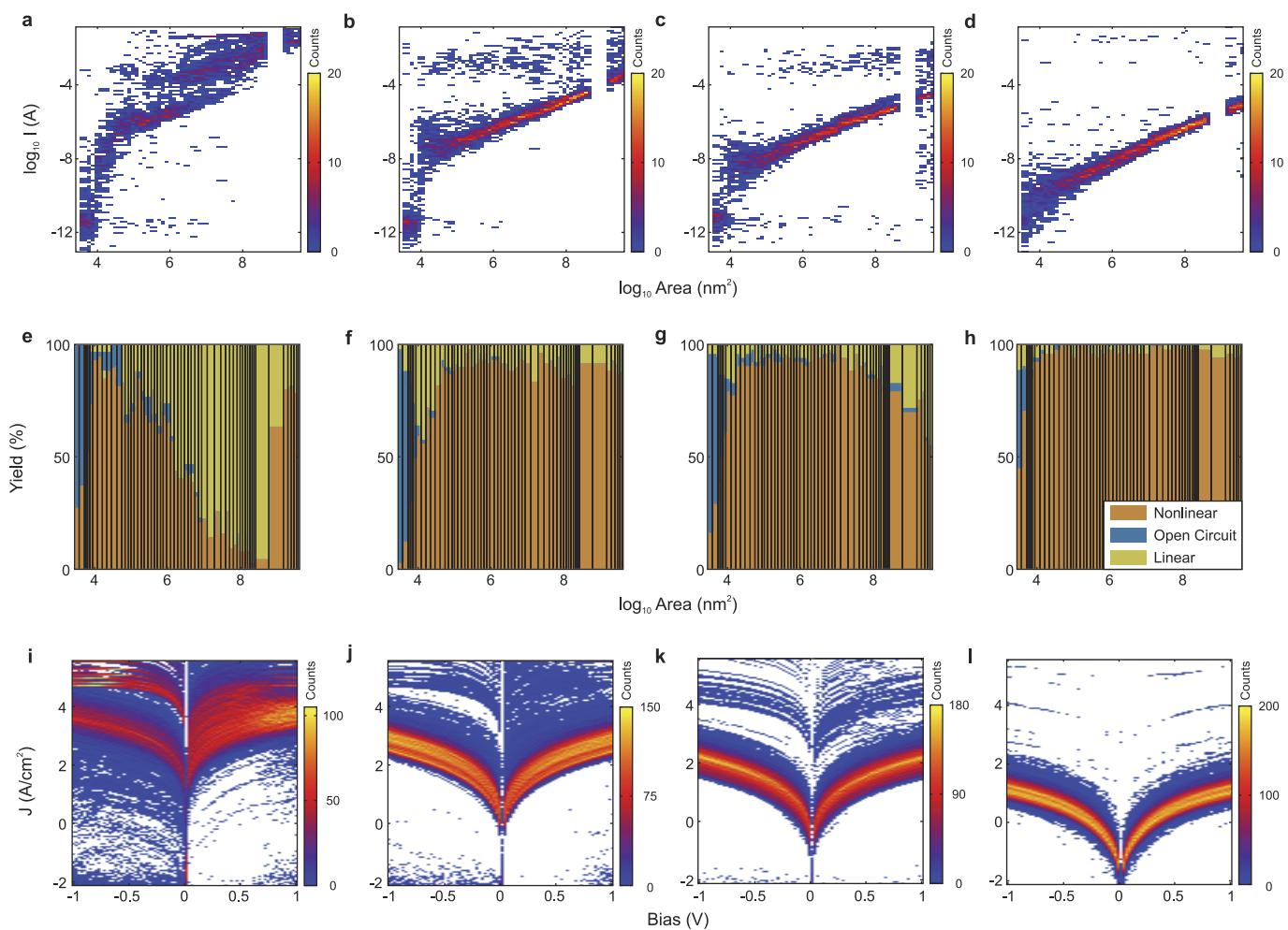
temperature dependence. **b, c**, Preliminary results of current density for C8 and C10 monolayers with two different nanoparticle diameters: 3 nm (**b**) and 5 nm (**c**). The current density decreases for larger nanoparticles.



Extended Data Fig. 6 | Short- and long-term stability. **a**, Histogram of the current ratio between upward and downward sweeps at 0.5 V, with no substantial change between the two sweeps. **b**, Histogram of the current ratio between the initial sweep and the sweep taken 136 days later, at 0.5 V. A large part of the devices did not change, or changed by only a small amount.



Extended Data Fig. 7 | Device scatter versus area. A comparison of the full width half maximum (FWHM) current density deviation σ_j against active device area for the literature quoted in Extended Data Table 1. We obtain similar or lower deviation than the literature samples over a considerably increased range in device area.



Extended Data Fig. 8 | Additional data for C4, C6, C8 and C10 devices. **a–d**, Histograms of I versus A . **e–h**, Device categorization for different values of A . **i–l**, Histograms of J versus A . The molecular length increases from left to right (**a, e, i**, C4; **b, f, j**, C6; **c, g, k**, C8; **d, h, l**, C10).

Extended Data Table 1 | Device properties in literature

Work	Method	J (A/cm ²)	FWHM σ _J log ₁₀ (A/cm ²)	# Devices measured	Overall Yield	Pore Diameter Range	Mass fabrication compatibility
Akkerman ⁷	PEDOT	2 10 ²	0.1-0.2	> 17	> 95 %	1-100 μm	limited
Neuhausen ¹⁵	Aedetron	5 10 ¹	0.4	> 13	70-100 %	0.3 / 1 μm	limited
Park ^{35,*}	PEDOT	3 10 ¹	0.3	> 20	-	1x2 mm ²	limited
Wang ³³	Direct Evaporation	4.9 10 ⁵	0.3	4800	1.75 %	2 μm	yes
Lee ^{34,*}	Direct Evaporation	8.8 10 ⁴	0.2	-	5 %	50 nm	yes
Loo ⁹	Transfer Printing	-	0.8	88	-	50 - 500 μm	limited
Jeong ^{38,*}	Transfer Printing	1.6 10 ²	1	128	70 %	2-5 μm	limited
Reus ^{37,*}	Direct eGaN	1.2 10 ⁻²	0.6-0.9	376	-	60 μm	no
Wang ⁸	Graphene	8 10 ⁴	0.8	~180	90 %	4 μm	limited
Seo ³⁶	rGO	2 10 ¹	0.1	50	> 99 %	100x100 μm ²	limited
This work	Nanoparticles	1.2 10 ² (5 10 ³)	0.9 overall 0.3 (10 μm pore)	3355	95 %	60 nm - 70 μm	yes

A comparison of current density, scatter, number of devices measured, overall yield, pore diameter variation and mass-fabrication compatibility for previous ensemble devices reported in literature^{7-9,15,33-38} and the current devices.

*Indicates use of C8 alkanemonothiols (instead of C8 alkanedithiols).

Electrically controlled water permeation through graphene oxide membranes

K.-G. Zhou^{1,2*}, K. S. Vasu^{1,2*}, C. T. Cherian^{1,2}, M. Neek-Amal^{3,4}, J. C. Zhang⁵, H. Ghorbanfekr-Kalashami⁴, K. Huang^{1,2}, O. P. Marshall⁶, V. G. Kravets⁶, J. Abraham^{1,2}, Y. Su^{1,2}, A. N. Grigorenko⁶, A. Pratt⁵, A. K. Geim⁶, F. M. Peeters⁴, K. S. Novoselov⁶ & R. R. Nair^{1,2*}

Controlled transport of water molecules through membranes and capillaries is important in areas as diverse as water purification and healthcare technologies^{1–7}. Previous attempts to control water permeation through membranes (mainly polymeric ones) have concentrated on modulating the structure of the membrane and the physicochemical properties of its surface by varying the pH, temperature or ionic strength^{3,8}. Electrical control over water transport is an attractive alternative; however, theory and simulations^{9–14} have often yielded conflicting results, from freezing of water molecules to melting of ice^{14–16} under an applied electric field. Here we report electrically controlled water permeation through micrometre-thick graphene oxide membranes^{17–21}. Such membranes have previously been shown to exhibit ultrafast permeation of water^{17,22} and molecular sieving properties^{18,21}, with the potential for industrial-scale production. To achieve electrical control over water permeation, we create conductive filaments in the graphene oxide membranes via controllable electrical breakdown. The electric field that concentrates around these current-carrying filaments ionizes water molecules inside graphene capillaries within the graphene oxide membranes, which impedes water transport. We thus demonstrate precise control of water permeation, from ultrafast permeation to complete blocking. Our work opens up an avenue for developing smart membrane technologies for artificial biological systems, tissue engineering and filtration.

Our devices are essentially graphene oxide membranes with metal electrodes on both sides, fabricated by depositing a thin (approximately 10 nm) gold (Au) film on top of the graphene oxide membrane, which is prepared on a porous silver (Ag) substrate (Extended Data Fig. 1). Such a thin layer of gold is sufficiently porous that it does not noticeably change the ultrafast water permeation (slip-enhanced flow^{17,22}) properties of the graphene oxide membranes (Extended Data Fig. 1d). For details of device fabrication and permeation experiments, see Methods sections ‘Fabrication of metal–graphene oxide–metal membranes’ and ‘Permeation experiments’. In Fig. 1a, b we show a schematic and an optical photograph of our membrane device.

To increase the electric field applied up to the values sufficient for water dissociation, thin conductive filaments were introduced in our membranes by controllable electrical breakdown. We attribute this conductive filament formation to the presence of moisture in graphene oxide membranes, which are known to facilitate the formation of conducting paths (generally carbon) across insulators under large electric fields²³ (see Methods section ‘Conducting filament formation’). In Fig. 1b we show the current–voltage (*I*–*V*) characteristics of the device exposed to 100% relative humidity during filament formation. Up to a critical voltage V_c , the current does not change appreciably. However, at V_c (roughly 2 V for the sample in Fig. 1 and varied by 25% for the four different samples studied), a partial electrical breakdown occurred, evident through a sudden increase in current up to the compliance level. This breakdown state is stable and characterized by low transversal resistance.

After the controllable breakdown, the *I*–*V* characteristics are nearly ohmic and suggest the existence of permanent electrically conductive channels (Fig. 1b). However, the in-plane *I*–*V* measurements do not reveal any substantial change in conductivity compared with pristine samples (Extended Data Fig. 2b). Unlike in-plane conductivity, the out-of-plane conductivity was found to be stable and insensitive to the humidity of the environment (Extended Data Fig. 2b). This confirms the formation of conducting filaments (such as carbon filaments) between the electrodes (out-of-plane) that are not connected in the plane of the membrane. Further, our peak-force tunnelling atomic force microscopy (PF TUNA) and Raman spectroscopy (see Methods section ‘Characterization of conducting filaments’) experiments clearly show the presence of conductive filaments with diameters of less than 50 nm (Fig. 1d, Extended Data Figs. 3, 4), with a filament density of about 10^7 cm^{-2} . No substantial changes were observed in the chemical stoichiometry of the graphene oxide membrane (as measured by X-ray photoelectron spectroscopy (XPS)), except for a small increase in the C/O ratio from 3.2 to 3.6 on the membrane surface closest to the positive electrode. The C/O ratio of all other membrane surfaces remained the same as the pristine sample (see Methods section ‘XPS’, Extended Data Fig. 5).

To probe the influence of electric field on water permeation through pristine graphene oxide membranes, the applied voltage was increased in a stepwise manner, which allowed us to monitor current and water permeation as functions of time during the filament formation process. At each voltage step, measurements were carried out for a minimum period of four hours. In Fig. 1e we show the weight loss of the sealed container and the corresponding water permeation rate during the filament formation process. We found no appreciable change in water permeation rate up to V_c (Fig. 1e). At V_c , we observed a sudden decrease (by a factor of 15) in water permeation after the partial breakdown of the graphene oxide membrane. Hereafter, water permeation through the membranes with low transverse resistance showed a strong dependence on the applied voltage, decreasing with increased voltage. At zero voltage, the water permeation recovered to almost the initial value of the pristine sample (about 85%). The water-vapour permeance of graphene oxide membrane that corresponds to this permeation rate is $6.5 \text{ l h}^{-1} \text{ m}^{-2} \text{ bar}^{-1}$, roughly twice that of commercial polyamide thin-film membranes with 10-times-smaller thickness (Extended Data Fig. 1d).

The stable out-of-plane electrical conductivity of the membrane and the electrical control of water permeation are more evident from the continuous forward and backward voltage sweeps performed after the filament formation (Fig. 1f). The electrically controlled water permeation was found to be independent of the polarity of the applied voltage and completely reversible, even after multiple voltage cycles switching between 0 and 1.8 V (Fig. 1f, left inset) and long-term switching off (Fig. 1f, right inset). The electrical control of water permeation was further confirmed by mass spectrometry (see Methods section

¹National Graphene Institute, University of Manchester, Manchester, UK. ²School of Chemical Engineering and Analytical Science, University of Manchester, Manchester, UK. ³Department of Physics, Shahid Rajaee Teacher Training University, Tehran, Iran. ⁴Department of Physics, University of Antwerp, Antwerp, Belgium. ⁵Department of Physics, University of York, York, UK.

⁶School of Physics and Astronomy, University of Manchester, Manchester, UK. *e-mail: kai-ge.zhou@manchester.ac.uk; vasusiddeswara.kalangi@manchester.ac.uk; rahul@manchester.ac.uk

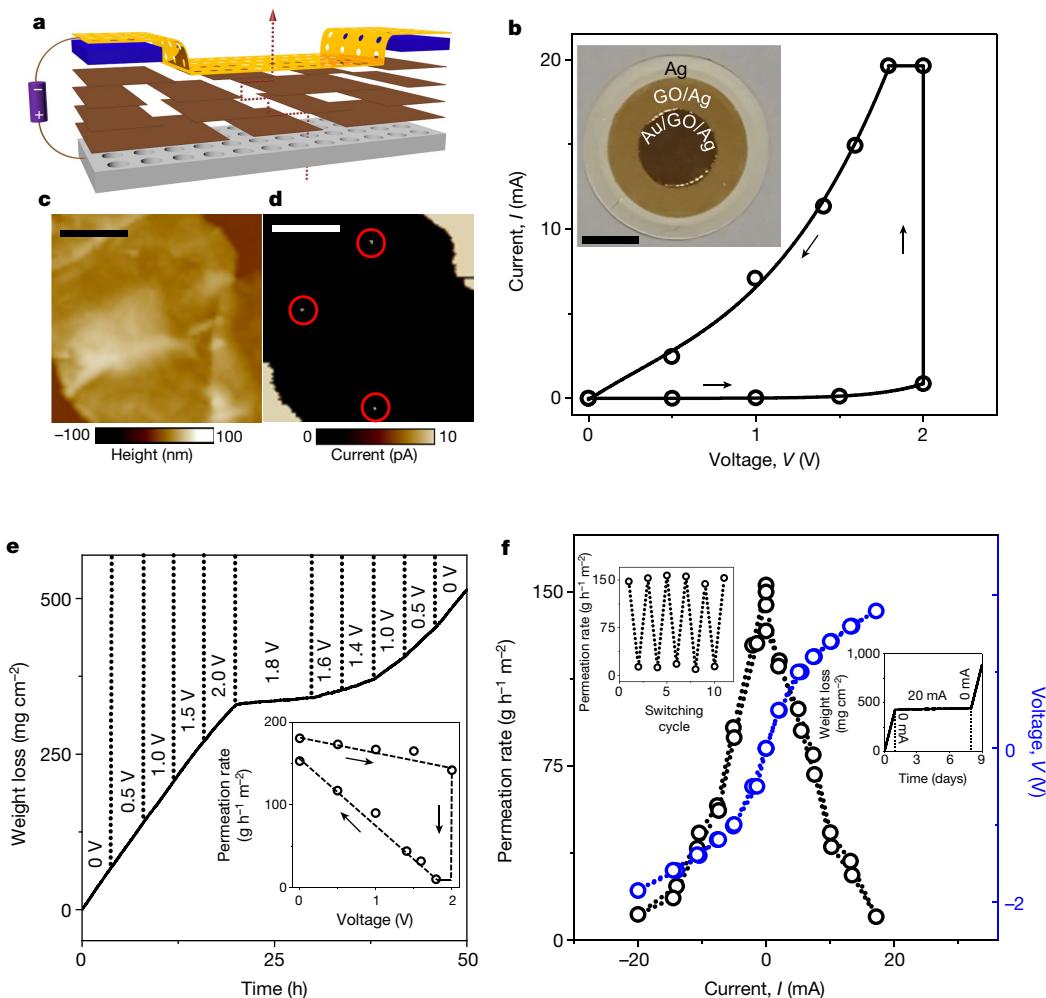


Fig. 1 | Electrically controlled water permeation through a graphene oxide membrane. **a**, Schematic of a graphene oxide membrane (deposited on a porous silver substrate) with voltage applied. The yellow, blue and brown layers represent the porous gold electrode, a plastic mask (that avoids electric shorting between the gold and silver electrodes) and graphene oxide sheets, respectively. The dotted line shows a possible pathway for water permeation. **b**, I – V characteristics at 100% relative humidity during the first voltage sweep show a sudden increase in current, suggesting partial electrical breakdown of the graphene oxide membrane and conducting filament formation. The solid line is a guide to the eye and arrows indicate the direction of the voltage sweep. Inset, photograph of a graphene oxide membrane showing the central sandwiched (Au/graphene oxide (GO)/Ag) region, the graphene oxide on supporting silver (GO/Ag) and the bare silver (Ag) substrate. The plastic mask outside the gold layer is removed for clarity. The outer area of the central sandwiched region was masked to block water permeation through this region (Extended Data Fig. 1a, b). Scale bar, 5 mm. **c, d**, Topography (c) and the corresponding PF TUNA current image (d) of a graphene oxide membrane after filament formation exfoliated on a gold-thin-film-coated silicon substrate.

‘Mass spectrometry’ (Extended Data Fig. 6). The electrical control that we observed can be extended to the case of liquid water permeation (see Methods section ‘Electrical control of liquid water permeation’, Extended Data Fig. 7).

To understand the influence of voltage and current on water permeation, we performed additional experiments using graphene oxide membranes with different permeation areas and thicknesses. In Fig. 2a we show the water permeation rate as a function of current through graphene oxide membranes with 1- μm thickness but two different permeation areas. Here, the smaller (by a factor of four) membrane requires only one-quarter of the current compared with the large membrane to achieve the same level of water blockage, suggesting

The conducting filaments (size ranging from 20 nm to 45 nm) formed in the graphene oxide membrane are marked by red circles. Scale bars, 1 μm . **e**, Weight loss for a water-filled container sealed with a 1- μm -thick graphene oxide membrane (7-mm diameter), at different voltages applied across the membrane during the filament formation process (as labelled). The vertical dotted lines show the times at which voltage switches. Inset, water permeation rate through the graphene oxide membrane as a function of applied voltage. The dashed line is a guide to the eye and arrows indicate the direction of the voltage sweep. **f**, Water permeation rate as a function of the current across the graphene oxide membrane after filament formation and the corresponding I – V characteristics (colour-coded axes). One complete voltage sweep for positive and negative polarity is plotted. Left inset, continuous switching between 0 and 1.8 V showing the stability of permeation control. Right inset, weight-loss data showing the stability of long-term switching off (seven days) and the subsequent recovery. The vertical dotted lines show the time at which the current switches. All weight-loss measurements were performed inside a dry chamber with 10% relative humidity.

that the current density through the membrane is the crucial factor in controlling water permeation. In Fig. 2b we show the normalized water permeation rate as a function of current through 1- μm -thick and 5- μm -thick graphene oxide membranes with the same permeation area. Here, the decrease in normalized water permeation rate was found to be nearly identical for the same magnitude of electric current passing through the membranes. However, the voltage required to maintain the same magnitude of current is very different, indicating a direct correlation between the water permeation rate and the electric current rather than the applied voltage. Both of these observations confirm that the current controls water transport in our experiments rather than the applied voltage.

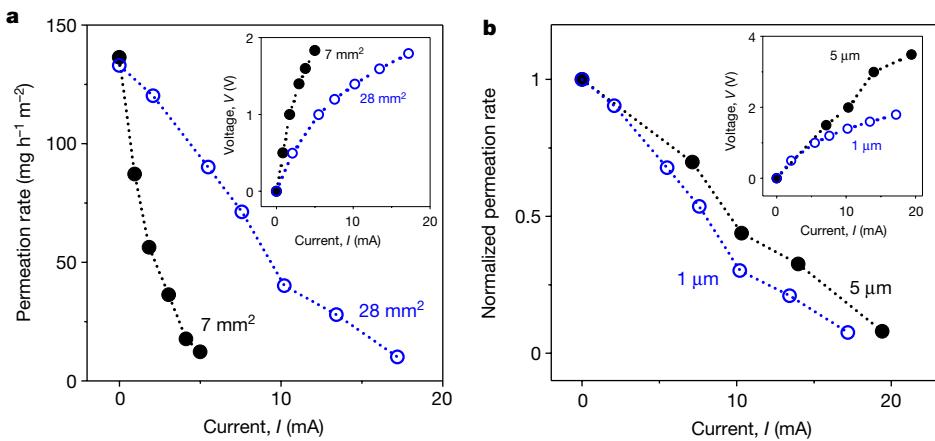


Fig. 2 | Current controlled permeation.

a, Permeation rate as a function of the current through two graphene oxide membranes with different areas (28 mm^2 , open blue circles; 7 mm^2 , filled black circles). **b**, Normalized water permeation rate as a function of the current through graphene oxide membranes with two different thicknesses. Permeation rates were normalized with respect to zero applied voltage because the absolute water permeation rates of $1\text{-}\mu\text{m}$ -thick (open blue circles) and $5\text{-}\mu\text{m}$ -thick (filled black circles) graphene oxide membranes were different. Insets in **a** and **b** show the corresponding I – V characteristics.

The influence of the current on the water transport can be due to Joule heating, dewetting or possible electrochemical changes. We measured the membrane temperature as a function of the electric current

across the membrane (see Methods section ‘Joule heating effect’) and found no substantial variations (Extended Data Fig. 8a). Further, we ruled out the prospect of changes in the wetting properties of graphene

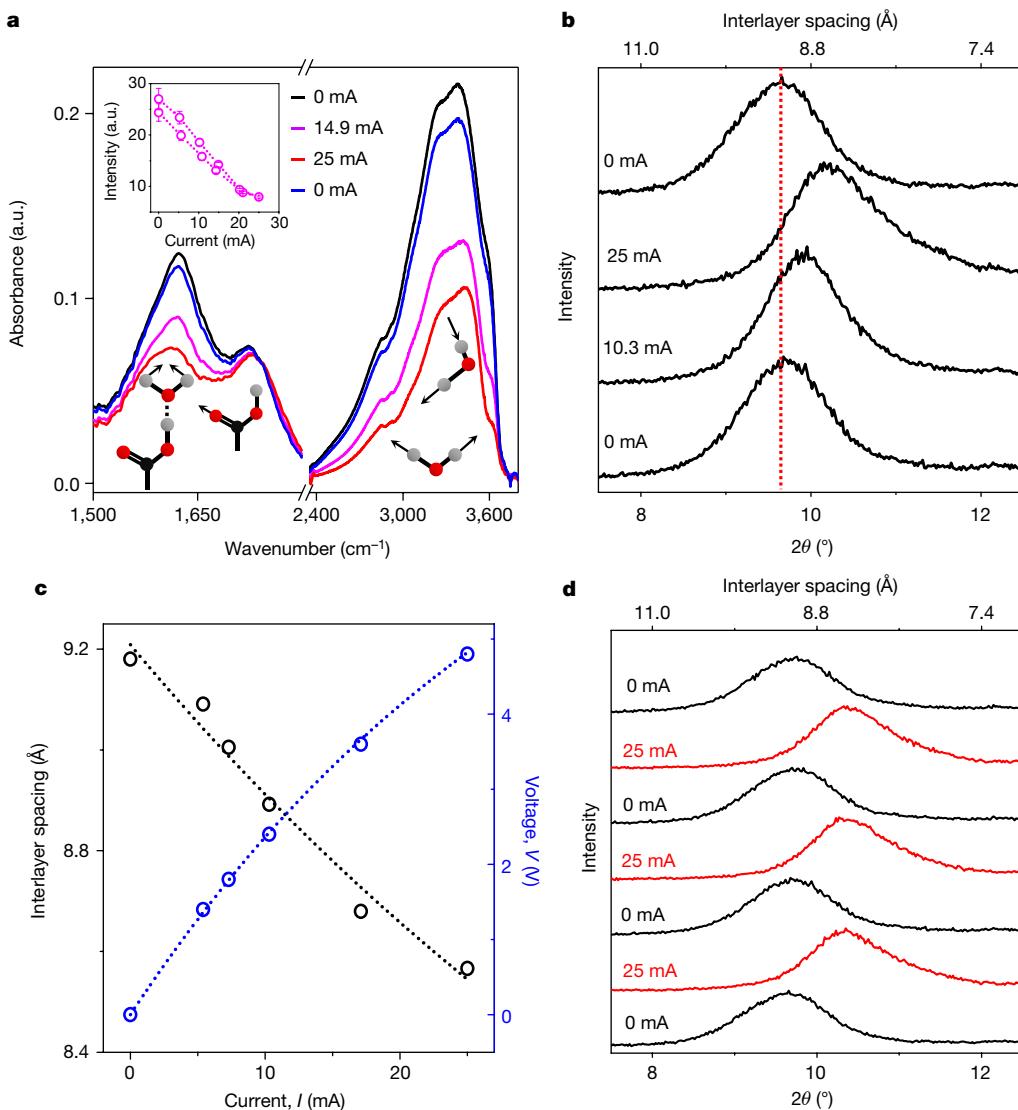


Fig. 3 | In situ Fourier-transform infrared and X-ray measurements.

a, In situ Fourier-transform infrared spectrum acquired from a graphene oxide membrane in which the current was cycled as $0 \text{ mA} \rightarrow 14.9 \text{ mA} \rightarrow 25 \text{ mA} \rightarrow 0 \text{ mA}$. Inset, peak intensity (area under the curve) as a function of the current for the peak at about $3,500 \text{ cm}^{-1}$ in forward and backward sweeps. The ball-and-stick models below the data illustrate the vibrational modes responsible for the different

infrared peaks. Red, oxygen; black, carbon; grey, hydrogen. **b**, XRD for different current levels across the $10\text{-}\mu\text{m}$ -thick graphene oxide membrane (membrane diameter of about 10 mm). The vertical dotted line shows the peak position for the zero current. **c**, Changes in the interlayer spacing as a function of current and the corresponding I – V behaviour of the membrane (colour-coded axes). The dotted lines are guides to the eye. **d**, Reversible switching of the (001) X-ray peak with the current.

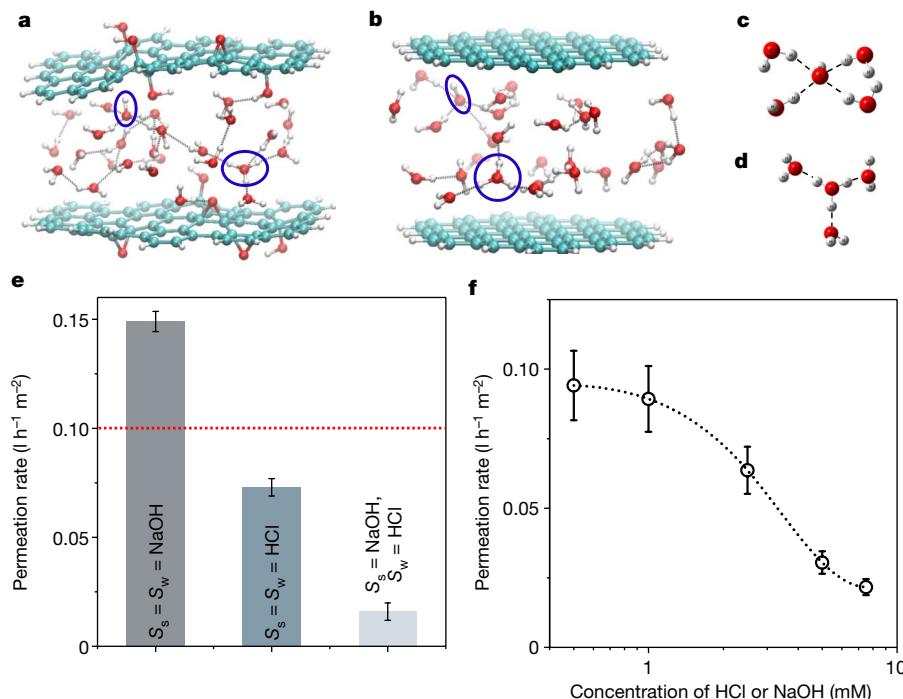


Fig. 4 | Influence of H_3O^+ and OH^- ions on water dynamics inside nanochannels. **a, b**, The first-principles optimized configuration for confined water in the presence of H_3O^+ and OH^- ions inside graphene oxide (**a**) and graphene (**b**) nanochannels with a height of 1 nm. White, cyan and red spheres represent hydrogen, carbon and oxygen atoms, respectively. The black dashed lines represent hydrogen bonds. Circles show H_3O^+ and OH^- ions. **c, d**, First hydration shell of OH^- (**c**) and H_3O^+ (**d**) ions inside the nanochannel. **e**, Osmotic-pressure-driven liquid water permeation rate in the presence of NaOH and/or HCl in water and

sucrose compartments separated by a graphene oxide membrane. NaOH or HCl solutions (10 mM) added to water and sucrose compartments are denoted as S_w and S_s , respectively. The red dotted line indicates the water permeation rate of pure water, that is, when NaOH and HCl are absent. **f**, Water permeation rate as a function of the concentration of NaOH and HCl in the sucrose and water compartments, respectively. The dotted line is a guide to the eye. Error bars in **e** and **f**, standard deviations from three different measurements using different membranes.

oxide: *in situ* water-absorption experiments confirm the absence of substantial changes in weight intake at 100% relative humidity, in the presence and absence of electric current (dewetting could cause a decrease in weight; see Methods section ‘*In situ* water absorption and release’, Extended Data Fig. 8b). To test the electrochemical mechanism, we measured the infrared modes of water from the graphene oxide membrane for a varying electric current across it (Fig. 3a; see Methods section ‘*In situ* infrared measurements’). The pristine sample has three main characteristic infrared bands, peaked at about $1,620 \text{ cm}^{-1}$, $1,737 \text{ cm}^{-1}$ and $3,500 \text{ cm}^{-1}$, which correspond to the deformation vibration of adsorbed water molecules, the carbonyl ($\text{C}=\text{O}$) stretching mode of the carboxylic group and the $\text{O}-\text{H}$ stretching mode in both the graphene oxide sheets and the interspersed water molecules, respectively^{24–26}. Interestingly, all the band intensities associated with water molecules decreased when an electric current was switched on and fully recovered to their initial values (Fig. 3a, inset) when the current was brought to zero. However, the band intensity relating to the carbonyl groups remained constant.

To test whether the decrease in the infrared water signal is associated with the reduced amount of water, we performed an X-ray diffraction (XRD) study in the presence of electric current in our membrane devices²¹ (see Methods section ‘*In situ* XRD measurements’). In Fig. 3b we show the *in situ* changes of the (001) reflection as a function of the current across the membrane, with the (001) peak clearly shifting to higher 2θ values at elevated currents. We estimated the interlayer spacing d from the XRD analysis and plotted it as a function of electric current (Fig. 3c). We found a decrease in d from 9.2 \AA to 8.5 \AA as the current across the membrane increased from 0 to 25 mA. These electric-current-induced changes in d were also found to be reversible (Fig. 3d) and did not present during the first voltage sweep up to V_c , before the filament formation. However, the small change in d that we observed (0.7 \AA) is not expected to affect the water permeation in

graphene oxide membranes substantially, owing to the slip-enhanced water permeation through graphene capillaries^{17,22}, and cannot explain the observed reduction (nearly 50%) in the infrared peak intensity.

On the basis of these observations, we attribute the electrically controlled water permeation to current-mediated ionization of water molecules. It is known that a current-carrying conductor produces an electric field around it^{27,28} (see Methods section ‘Electric field due to a current-carrying conductor’). The exact value of the electric field depends on the parameters of the set-up, but for a coaxial arrangement (current flows in one direction through the inner conductor of radius a and in the other direction through a coaxial conductor of radius b —in our case the characteristic size of the sample) the radial component of the field is

$$E_r = \frac{Jz}{\sigma r} \frac{1}{\ln(a/b)}$$

where J is the current density, z varies from 0 to L (the length of the wire or thickness of the membrane), r is the radial distance from the centre of the wire and σ is its electrical conductivity^{27,28} (Extended Data Fig. 9a). In our case, this formula reduces to $E_r = (V/r)/\ln(a/b)$ and it is obvious that, close to the surface of the filaments (where r is tens of nanometres), the electric field can be as high as around 10^7 V m^{-1} (Extended Data Fig. 9b). Such large electric fields could dissociate water molecules to produce hydronium (H_3O^+) and hydroxyl (OH^-) ions (see Methods section ‘Electric-field-induced dissociation of water’), with the effect becoming more pronounced at higher currents.

Further, molecular dynamics simulations performed to understand the influence of H_3O^+ and OH^- ions on water permeation show that the water permeation rate in graphene capillaries decreases with increasing ion concentration (see Methods section ‘Molecular dynamics simulations’, Extended Data Fig. 10). This could be attributed to two main effects: (i) the localization of H_3O^+ and OH^- ions inside the

capillary, owing to their interaction with graphene or graphene oxide, which creates a local blockage; or (ii) the formation of large hydrated clusters due to strong interactions between H_3O^+ or OH^- ions and surrounding water molecules, which impedes the water transport. Our first-principles calculations²⁹ confirm that the latter is the dominant effect, because of the strong hydrogen bond between the ions and surrounding water (Fig. 4a–d; see Methods section ‘Interaction of H_3O^+ and OH^- ions with water and graphene or graphene oxide capillaries’). In agreement with this modelling result, additional liquid water permeation experiments clearly demonstrate that the simultaneous presence of both H_3O^+ and OH^- ions are necessary to reduce the water permeation rate (Fig. 4e, f; see Methods section ‘Influence of H_3O^+ and OH^- ions’). This is consistent with the previously reported cooperativity effect³⁰, whereby the simultaneous presence of strongly hydrated cations and anions decreases the mobility of water molecules and changes their dynamics. Our proposed model of electric-field-enhanced dissociation of water is also consistent with the reversible chemical changes in the interlayer water molecules (infrared peak intensities) that we observed and the changes in the interlayer spacing (Fig. 2b) with current. These changes could be attributed to the changes in the structure of interlayer water due to the ionization. Further experimental and theoretical efforts are needed for a comprehensive understanding of the exact mechanism responsible for electric-current control of water permeation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0292-y>.

Received: 22 October 2017; Accepted: 14 May 2018;

Published online 11 July 2018.

1. Karnik, R. et al. Electrostatic control of ions and molecules in nanofluidic transistors. *Nano Lett.* **5**, 943–948 (2005).
2. Gravelle, S. et al. Optimizing water permeability through the hourglass shape of aquaporins. *Proc. Natl Acad. Sci. USA* **110**, 16367–16372 (2013).
3. Liu, Z., Wang, W., Xie, R., Ju, X.-J. & Chu, L.-Y. Stimuli-responsive smart gating membranes. *Chem. Soc. Rev.* **45**, 460–475 (2016).
4. Xiao, K. et al. Electrostatic-charge- and electric-field-induced smart gating for water transportation. *ACS Nano* **10**, 9703–9709 (2016).
5. Wang, Z. et al. Polarity-dependent electrochemically controlled transport of water through carbon nanotube membranes. *Nano Lett.* **7**, 697–702 (2007).
6. Hetherington, A. M. & Woodward, F. I. The role of stomata in sensing and driving environmental change. *Nature* **424**, 901–908 (2003).
7. Borgnia, M. J., Nielsen, S., Engel, A. & Agre, P. Cellular and molecular biology of the aquaporin water channels. *Annu. Rev. Biochem.* **68**, 425–458 (1999).
8. Zhao, C., Nie, S., Tang, M. & Sun, S. Polymeric pH-sensitive membranes—a review. *Prog. Polym. Sci.* **36**, 1499–1520 (2011).
9. Kou, J. et al. Electromanipulating water flow in nanochannels. *Angew. Chem. Int. Ed.* **54**, 2351–2355 (2015).
10. Li, J. et al. Electrostatic gating of a nanometer water channel. *Proc. Natl Acad. Sci. USA* **104**, 3687–3692 (2007).
11. Gong, X. et al. A charge-driven molecular water pump. *Nat. Nanotechnol.* **2**, 709–712 (2007).
12. Vaitheswaran, S., Rasaiah, J. C. & Hummer, G. Electric field and temperature effects on water in the narrow nonpolar pores of carbon nanotubes. *J. Chem. Phys.* **121**, 7955–7965 (2004).
13. Saitta, A. M., Saija, F. & Giacinta, P. V. Ab initio molecular dynamics study of dissociation of water under an electric field. *Phys. Rev. Lett.* **108**, 207801 (2012).
14. Qiu, H. & Guo, W. Electromelting of confined monolayer ice. *Phys. Rev. Lett.* **110**, 195701 (2013).

15. Choi, E.-M., Yoon, Y.-H., Lee, S. & Kang, H. Freezing transition of interfacial water at room temperature under electric fields. *Phys. Rev. Lett.* **95**, 085701 (2005).
16. Diallo, S. O., Mamontov, E., Nobuo, W., Inagaki, S. & Fukushima, Y. Enhanced translational diffusion of confined water under electric field. *Phys. Rev. E* **86**, 021506 (2012).
17. Nair, R. R., Wu, H. A., Jayaram, P. N., Grigorieva, I. V. & Geim, A. K. Unimpeded permeation of water through helium-leak-tight graphene-based membranes. *Science* **335**, 442–444 (2012).
18. Joshi, R. K. et al. Precise and ultrafast molecular sieving through graphene oxide membranes. *Science* **343**, 752–754 (2014).
19. Sun, P., Wang, K. & Zhu, H. Recent developments in graphene-based membranes: structure, mass-transport mechanism and potential applications. *Adv. Mater.* **28**, 2287–2310 (2016).
20. Liu, G., Jin, W. & Xu, N. Graphene-based membranes. *Chem. Soc. Rev.* **44**, 5016–5030 (2015).
21. Abraham, J. et al. Tunable sieving of ions using graphene oxide membranes. *Nat. Nanotechnol.* **12**, 546–550 (2017).
22. Radha, B. et al. Molecular transport through capillaries made with atomic-scale precision. *Nature* **538**, 222–225 (2016).
23. Kao, K.-C. *Dielectric Phenomena in Solids: With Emphasis on Physical Concepts of Electronic Processes* Ch. 8 (Academic Press, Amsterdam, 2004).
24. Acik, M. et al. The role of oxygen during thermal reduction of graphene oxide studied by infrared absorption spectroscopy. *J. Phys. Chem. C* **115**, 19761–19781 (2011).
25. Hontoria-Lucas, C., López-Peinado, A. J., López-González, J. D., Rojas-Cervantes, M. L. & Martín-Aranda, R. M. Study of oxygen-containing groups in a series of graphite oxides: physical and chemical characterization. *Carbon* **33**, 1585–1592 (1995).
26. Konkena, B. & Vasudevan, S. Understanding aqueous dispersibility of graphene oxide and reduced graphene oxide through pKa measurements. *J. Phys. Chem. Lett.* **3**, 867–872 (2012).
27. Jackson, J. D. Surface charges on circuit wires and resistors play three roles. *Am. J. Phys.* **64**, 855–870 (1996).
28. Marcus, A. The electric field associated with a steady current in long cylindrical conductor. *Am. J. Phys.* **9**, 225–226 (1941).
29. Chen, L. et al. Ion sieving in graphene oxide membranes via cationic control of interlayer spacing. *Nature* **550**, 380–383 (2017).
30. Tielrooij, K. J., Garcia-Araez, N., Bonn, M. & Bakker, H. J. Cooperativity in ion hydration. *Science* **328**, 1006–1009 (2010).

Acknowledgements This work was supported by the Royal Society, Engineering and Physical Sciences Research Council, UK (EP/K016946/1, EP/N013670/1 and EP/P00119X/1), British Council (award reference number 279336045), European Research Council (contract 679689) and Lloyd’s Register Foundation. We thank J. Waters for assisting with X-ray measurements and G. Yu for electrical measurements.

Reviewer information *Nature* thanks H. Fang, N. Koratkar, B. Mi and H. B. Park for their contribution to the peer review of this work.

Author contributions R.R.N. initiated and supervised the project. K.-G.Z. performed the experiment and analysed the data with help from K.S.V. and R.R.N.; K.S.V. carried out the PF TUNA and Raman characterization and analysis. C.T.C. carried out the mass spectroscopy. K.H., J.A. and Y.S. helped in sample preparation, characterization and data analysis. K.S.V., M.N.-A., H.G.-K., K.S.N. and F.M.P. performed the theoretical modelling and simulations. J.C.Z. and A.P. performed the XPS characterizations. O.P.M., V.G.K. and A.N.G. performed the infrared characterizations. A.K.G. contributed to theoretical discussions. R.R.N., K.S.V., K.-G.Z. and K.S.N. co-wrote the paper. All authors contributed to discussions.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0292-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.R.N., K.-G.Z. or K.S.V.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Fabrication of metal–graphene oxide–metal membranes. Graphene oxide (GO) aqueous dispersions (flake size of approximately 10 μm) were prepared by exfoliation of graphite oxide powder (BGTE Materials) in water using bath sonication^{17,21}. Different steps in the fabrication process and an optical photograph of the metal–GO–metal sandwich membrane are shown in Extended Data Fig. 1a, b. First, a Sterlitech porous silver metal membrane (0.2 μm pore size and 13 mm diameter), which also acts as the bottom electrode, was used in a standard vacuum filtration set-up to prepare GO membranes for the permeation experiments. These membranes were glued (using Stycast 1266 epoxy resin) onto the polyethylene terephthalate (PET) films (step 1 in Extended Data Fig. 1a) with a circular aperture (diameter of about 1 cm) before depositing a thin (around 10 nm) top Au electrode (step 2 in Extended Data Fig. 1a) using thermal evaporation at a deposition rate of 0.8 Å s^{-1} in a high-vacuum chamber (2×10^{-6} mbar) at temperature $20 \pm 1^\circ\text{C}$. Here, the PET film prevents electrical shorting between the top and bottom electrodes in the metal–GO–metal sandwich membrane structures. Finally, these structures were glued onto another plastic disk (step 3 in Extended Data Fig. 1a), which provides mechanical support for sealing the sample to a stainless-steel water container used in pervaporation experiments¹⁷.

Permeation experiments. We used a previously reported gravimetric method¹⁷ to probe the electrical effects on water permeation through GO membranes. A plastic disk containing the metal–GO–metal sandwich membrane was fixed to a stainless-steel water container using two rubber O-rings to ensure an air-tight seal. The water permeation rate was measured in terms of weight loss of the water container using a computer-controlled precision balance (Mettler Toledo; accuracy 0.1 mg). All the gravimetric experiments were carried out in a chamber with a controlled relative humidity (RH) of 10%.

To probe electrical effects on water permeation, thin copper wires extending from both the electrodes were connected to a Keithley 2410 sourcemeter (Extended Data Fig. 1a). An additional Keithley 2182A Nanovoltmeter was also connected across the membrane to measure the potential drop. During permeation experiments a current compliance (20 mA for the sample in Fig. 1) was set to limit the current density to a maximum of 70 mA cm^{-2} to prevent uncontrollable breakdown (blustering) of the device.

To study the influence of the Au electrode on the water permeation rate, we deposited thin Au layers (10–50 nm) on 1- μm -thick GO membranes. Extended Data Fig. 1c shows a scanning electron microscope (SEM) image of a 10-nm Au film on a GO membrane, which clearly displays different types of uniformly distributed pores (discontinuities and voids) occupying approximately 40% of the total area of the Au film. Despite uniform pore distribution, 10-nm Au thin films still have a conductivity of $3.7 \times 10^6 \Omega^{-1} \text{ m}^{-1}$ (roughly 9% of bulk conductivity)³¹, enabling their use as electrodes. These Au electrodes become continuous with decreasing porosity (Extended Data Fig. 1d inset) as their thickness increases.

As expected, the permeation rate of a 1- μm -thick GO membrane on a porous Ag support (before Au electrode deposition) is found to be approximately same as that of the bare porous Ag support¹⁷ (Extended Data Fig. 1d). The water permeation rate decreased by a factor of only about 1.5 after depositing an Au electrode with a thickness of about 10 nm on the GO membrane. However, with increasing thickness of the Au electrode, the water permeation rate of Au/GO/Ag membranes decreases exponentially (Extended Data Fig. 1d) owing to the decreased pore size of the electrode. The water permeation rate of Au/GO/Ag membranes with a 10-nm Au electrode is roughly two times greater than that of a commercial polyamide nanofiltration membrane (NF Polyamide-TFC, 200–400 Da, Dow Filmtec; Extended Data Fig. 1d).

Conducting filament formation. Controllable dielectrical breakdown of the GO membrane was performed by setting appropriate current compliance during the first voltage sweep across the membrane. Extended Data Fig. 2a shows I – V characteristics of the GO membranes at different RH in the first voltage sweep. Similar to the tracking phenomenon in dielectrics^{23,32,33}, we found that conducting filament formation in GO membranes is highly sensitive to the humidity of the environment. The sample exposed to zero humidity did not show any evidence of the formation of electrically conducting filaments even up to 50 V, confirming the high dielectric strength of GO in a dry atmosphere^{34,35}. However, the samples exposed to humid conditions deviated from this behaviour, showing a sudden increase in the current at V_c , where the samples were permanently switched to a conducting state with stable out-of-plane conductivity even in the negative bias conditions (Fig. 1f). The decrease in V_c with increasing humidity of the environment (Extended Data Fig. 2a) further suggests the contribution of absorbed water content (interlayer water) in the formation of the conducting filaments inside GO membranes.

Similar conducting filament formation has previously been observed in dried GO thin films (thicknesses of 10–100 nm) used for the fabrication of resistive random-access memory (ReRAM) devices. Nearly two orders of magnitude higher electric field across such thin films, compared with the field across our micro-metre-thick GO membranes, causes the filament formation in dry conditions.

Importantly, the filament formation in ReRAM devices is completely reversible owing to the current-induced heating effects^{36,37}, and the GO thin film becomes electrically insulating once the polarity of the applied voltage changes. However, the filament formation in GO membranes in this study is permanent and the membranes remain permanently conducting regardless of the changes in the polarity of the applied voltage (Fig. 1f). Micrometre-length filaments and high filament density (roughly 10^7 cm^{-2}) in our GO membranes limit the current flow (see below) through each filament to a small value (about 1 nA), and hence the membranes stay in a permanently conducting state owing to negligible current-induced heating effects.

To understand electrical conductivity in GO membranes after the filament formation further, we measured the in-plane conductivity of a GO membrane that displayed out-of-plane conductivity. For this, the membrane was peeled off from the silver substrate and then exfoliated using scotch tape. A pair of electrodes (3 mm apart with a width of 1 cm) was made on this freshly peeled GO membrane using conductive silver paste to measure the in-plane conductivity. Extended Data Fig. 2b compares the out-of-plane and in-plane I – V characteristics of the GO membrane after conducting filament formation. Surprisingly, we found that the in-plane electrical conductivity of these GO membranes is similar to that of highly resistive pristine GO membranes (about $15 \mu\text{S cm}^{-1}$)³⁸. In addition, after filament formation, the stable out-of-plane I – V characteristics of the GO membrane (Extended Data Fig. 2b inset) recorded at 0% RH and 100% RH show humidity-independent out-of-plane electrical conductivity, unlike the in-plane electrical conductivity in GO³⁹. This clearly confirms the permanent filament formation. Additional experiments carried out after exposing the GO membrane containing filaments to 30 bar external pressure using nitrogen gas in the high-pressure vessel did not show any change in out-of-plane electrical conductivity or the electrical control on water permeation—establishing the stability of permanent conducting filaments.

Characterization of conducting filaments. To demonstrate the existence of conducting filaments inside the GO membranes, the top and bottom metal electrodes were peeled off using scotch tape. These GO membranes were further exfoliated using scotch tape to obtain the freshly cleaved inner surface for characterization using SEM, XPS, energy dispersive X-ray (EDX) and Raman spectroscopy. The SEM image of a conducting GO membrane shown in Extended Data Fig. 3a exhibits a texture similar to a pristine GO membrane with no noticeable features corresponding to the conducting filaments. EDX analysis also confirms the presence of only carbon and oxygen elements in the membrane.

Raman spectra of the GO samples were collected using HORIBA's XploRA PLUS Raman spectrometer with 1,800 lines mm^{-1} grating and 532-nm laser excitation at a power of 1.35 mW. Extended Data Fig. 3b, c shows the Raman I_D/I_G mapping for a pristine GO membrane and a GO membrane (close to the positive electrode) with conducting filaments (conducting GO membrane). I_D/I_G represents the ratio of intensities corresponding to the D band at $1,351 \text{ cm}^{-1}$ and the G band at $1,594 \text{ cm}^{-1}$ in the Raman spectrum (Extended Data Fig. 3e, f). As shown in Extended Data Fig. 3b, I_D/I_G is uniform (0.93) over the whole area of $10 \mu\text{m} \times 10 \mu\text{m}$ in the case of the pristine GO sample. However, for the GO membrane with conducting filaments, I_D/I_G was found to be inhomogeneous (Extended Data Fig. 3c), varying from 0.93 (blue areas) to 1.1 (green areas) at different positions on the sample. A similar increase in I_D/I_G was reported for the reduction of GO and attributed to an increase in the sp^2 carbon network⁴⁰. This suggests that the conducting filaments across the GO membrane are made of sp^2 carbon.

To investigate the uniformity of carbon filaments across the top and bottom electrodes, we carried out I_D/I_G mapping experiments on a conducting GO membrane surface close to the negative electrode, obtained by multiple peelings of the membrane. The I_D/I_G map acquired from the surface close to the negative electrode (Extended Data Fig. 3d) shows a substantial reduction in the size of domains corresponding to a high I_D/I_G ratio (about 1.1), suggesting the shrinkage of carbon filaments towards the negative electrode. This is further corroborated from the estimated sp^2 carbon content in XPS analysis (see Methods section 'XPS'). In addition, it is apparent from Extended Data Fig. 3c, d that several filaments are formed in each GO flake (roughly of size $10 \mu\text{m} \times 10 \mu\text{m}$). The number of conducting filaments estimated by counting bright spots in Extended Data Fig. 3d is about 10^7 cm^{-2} .

Further, we carried out atomic force microscopy (AFM) imaging using Bruker Dimension ICON AFM operated in PF TUNA mode (with Pt/Ir coated Bruker's PF TUNA probes) to confirm filament formation in conducting GO membranes. A pristine GO membrane was used as a reference. The surface close to the negative electrode of conducting GO membranes was exfoliated onto a Cr/Au (5 nm/95 nm) thin-film-deposited Si substrate to perform PF TUNA experiments. This facilitates the application of a constant DC bias (3–5 V with variable gain setting of 10^9 – 10^{10} V A^{-1}) between the sample and AFM probe. The current passing through the sample between the Cr/Au thin film and the AFM probe is measured using a current sensor, when the probe and sample are intermittently brought into contact. Thus, the mapping of electrical current across the samples provides a TUNA current image along with topography.

Extended Data Fig. 3g, i shows height images of the pristine and conducting GO (prepared at 100% RH) membranes exfoliated on the Cr/Au-deposited Si substrate. Their corresponding TUNA current images are shown in Extended Data Fig. 3h, j. There are no substantial differences in the topography of both the membranes in height images, and the measured thicknesses of the membranes varied from about 30 nm to 40 nm. However, the TUNA current images of pristine and conducting GO membranes reveal considerable differences. In the case of the pristine GO membrane, an apparent contrast was observed for GO (dark regions) from the surrounding gold thin film (bright region) because the magnitude of the TUNA current between the gold thin film and the probe is much larger than that between GO and the probe. On the other hand, TUNA current images (Fig. 1, Extended Data Fig. 3) obtained from conducting GO membranes (ten different areas of three samples) clearly display the presence of small conducting regions (which cannot be identified in height images) within non-conducting GO (dark) regions. Further, the filament density of about 10^7 cm^{-2} estimated by counting such conducting regions in the TUNA current image is in good agreement with Raman analysis.

On the basis of Raman and PF TUNA experiments, we propose the structure of conducting filaments as the parallel resistor model (Extended Data Fig. 3k). To validate this model, we divided a large conducting GO membrane (7-mm diameter) into four equal pieces. We found that the resistance of four individual pieces is relatively same in magnitude and four times greater than that of the large parent membrane (Extended Data Fig. 3k), but with similar resistivity. This clearly confirms the uniform distribution of conducting filaments inside the GO membrane, supporting the parallel-resistor model.

Additional PF TUNA images (Extended Data Fig. 4b, d) obtained for the GO membranes with filament formation occurring at 40% RH and inside liquid water (Extended Data Fig. 7) clearly confirm the influence of intercalated water on the conducting filament number density. Increased water content of GO membranes was found to increase the filament density and the diameter.

To understand filament formation further, we carried out additional experiments on partially reduced GO membranes (Extended Data Fig. 4e), prepared by annealing GO membranes (before top-electrode deposition) at 90 °C in an inert atmosphere. A GO membrane annealed at 90 °C showed a one-order increase in out-of-plane electrical conductivity, but no metallic behaviour, indicating only a partial reduction in oxygen content, whereas a GO membrane annealed at 120 °C became ohmic and did not exhibit filament formation.

I–V characteristics (Extended Data Fig. 4e) of a Au/partially reduced GO/Ag membrane during the first voltage sweep at 100% RH show much smaller V_c (less than 1 V) compared with pristine GO membranes. Interestingly, a PF TUNA current image of the partially reduced GO membrane after filament formation (Extended Data Fig. 4f) shows conducting filaments about three times larger in diameter, but with a conducting filament density ($2.4 \times 10^7 \text{ cm}^{-2}$) approximately equal to that of pristine GO membranes. The decreased V_c and formation of wider filaments in partially reduced GO membranes could be attributed to the increased conducting graphitic regions after partial reduction, consistent with a previously reported conducting-island-activated filament formation mechanism^{36,37}.

XPS. To investigate the chemical stoichiometry of GO membranes before and after the application of an electric potential (after electrically controlled water permeation experiments), we performed XPS experiments in an ultrahigh-vacuum system with a base pressure of less than 3×10^{-10} mbar using a monochromated Al K α source at 1,486.6 eV (Omicron XM 1000) and a power of 220 W.

Extended Data Fig. 5a shows XPS spectra from the pristine GO membrane and Extended Data Fig. 5b, c represents XPS spectra of a GO membrane after electrically controlled permeation, acquired from an inner surface of the membrane and a surface close to the positive electrode, respectively. The membrane surface close to the electrode was obtained by removing the Ag electrode through mechanical peeling, and then cleaving the membrane using scotch tape to reveal an inner surface. Using XPS Peak 4.1, each C 1s spectrum was fitted with four components representing the main bonding environments found in GO: C–C (284.5–284.8 eV), C–OH (285.2–285.4 eV), C–O–C (286.3–286.9 eV), and C=O and C(=O)–(OH) (287.8–289.1 eV)^{41,42}. We followed the widely accepted peak position assignment and the depiction of XPS peak fitting of GO^{41,42}, although different reports follow different representations. C/O ratios calculated from the fitted peak areas were found to be similar (3.2) for pristine GO and the inner surface of the membrane after electrically controlled permeation. By contrast, the membrane surface close to the positive electrode shows an increase in C/O ratio (3.6), indicating a higher sp^2 fraction close to the electrode. C–C fractions are found to be 56% and 63%, respectively, for the inner surface and the surface close to the electrode. This increase in sp^2 fraction close to the positive electrode is attributed to the formation of conducting carbon filaments after the application of a voltage across the GO membrane where the concentration of filaments is expected to be large.

Mass spectrometry. Electrical control of water permeation through GO membranes was also confirmed using mass spectrometry (MS). Here, a Au/GO/Ag sandwich structure was placed between two rubber O-rings in a custom-made

permeation cell (Extended Data Fig. 6a). Copper leads from the top and bottom electrodes were connected to a sourcemeter via an electrical feedthrough. Water vapour (25 mbar) and helium (25 mbar) were fed into the top chamber and permeation through the sample was monitored using mass spectrometry on the permeate side, maintaining 10^{-6} bar. We used a quadrupole residual gas analyser (HPR 30 Hiden Analytical) to measure the partial pressure of permeated species and wet cotton in the top chamber as a constant feed for water vapour. Extended Data Fig. 6b shows the partial pressure of water ($p_{\text{H}_2\text{O}}$), hydrogen, oxygen and helium (He) in the permeate side at different currents through the membrane. No appreciable change in helium partial pressure is observed during voltage cycling, confirming the membrane stability (no damage) under electric field and the impermeability of helium through conducting GO membranes¹⁷. Extended Data Fig. 6b, c shows that $p_{\text{H}_2\text{O}}$ decreases with increasing current through the membrane, suggesting a decrease in water permeation through the GO membrane with increasing current, which is consistent with the gravimetric measurements. In addition, the lack of any substantial change in H_2 or O_2 partial pressure at the permeate side indicates that H_2 and O_2 are not released during voltage cycling, ruling out the possibility of electrolysis of water. Extended Data Fig. 6b also confirms the reversible control of water permeation by electrical means.

Electrical control of liquid water permeation. We carried out osmotic-pressure-driven water permeation experiments to investigate the electrical control of liquid water permeation through Au/GO/Ag membranes. To this end, a Au/GO/Ag membrane was firmly fixed (Extended Data Fig. 7a) between two compartments (one filled with 12 ml of de-ionized water and other with 1 M sucrose solution) using Stycast 1266 epoxy resin. The liquid water permeation rate (flux) of the Au/GO/Ag membrane, calculated from the changes in volume of both the compartments (due to osmotic pressure) as a function of time, was found to be approximately $0.081 \text{ h}^{-1} \text{ m}^{-2}$, consistent with previous reports¹⁸. To introduce the conducting filaments in the GO membrane, controllable electrical breakdown of the Au/GO/Ag membrane was carried out while it was in contact with liquid water.

From the *I–V* characteristics shown in Extended Data Fig. 7b, we noticed that V_c for partial electrical breakdown was reduced to 0.8 V for the Au/GO/Ag membrane in contact with liquid water, very small compared with V_c for the sample at 100% RH. Extended Data Fig. 7c shows the water permeation rate as a function of current and the corresponding *I–V* characteristics. Similar to the water vapour permeation rate, the liquid water permeation rate also decreased with increasing electric current; however, only by a factor of two even for an electric current flow of 40 mA across the membrane. This smaller reduction in liquid water permeation rate compared with water vapour could be understood from the presence of about 25 bar osmotic pressure, which makes water flow faster even in the presence of H_3O^+ and OH^- ions, and the high number density of filaments (Extended Data Fig. 4d), which limits current flow through each filament to a smaller value, reducing the electric field strength required for dissociation of water.

We were limited to performing electrically controlled liquid water permeation experiments with applied voltages below 1 V. Because the membrane is in contact with liquid water, electrolysis of water takes place once the applied voltage is greater than 1 V and the H_2 and O_2 gases released at the electrodes damage the membrane.

Joule heating effect. To probe electric-current-induced Joule heating, we used an infrared thermometer (N92FX, Maplin) to monitor the variation in membrane temperature as a function of electric current across the membrane. No substantial changes were found in the membrane temperature for each current (Extended Data Fig. 8a). A temperature increase of only about 1 °C was measured for a current of 20 mA, which eliminates Joule heating effects in our water permeation experiments.

In situ water absorption and release. We performed in situ water absorption experiments on Au/GO/Ag membranes after filament formation to understand the changes in the wetting properties of GO while electric current flow is set across the membrane. After filament formation, the weight of the Au/GO/Ag membrane was measured continuously using a computer-controlled balance placed inside an environmental chamber at 35% and 100% RH in the absence and presence of electric current flow. The weight of a completely dried Au/GO/Ag membrane at 0% RH was used as a reference to calculate the weight intake of the membrane at different humidity and electric current.

The weight intake is increased from about 20% to 48% (Extended Data Fig. 8b) as the RH increases from 35% to 100% in the absence of current flow (0 mA) across the Au/GO/Ag membrane. We then set an electric current flow of 20 mA across the membrane and found no substantial change in weight intake (Extended Data Fig. 8b) within the experimental accuracy. This clearly suggests the absence of additional water absorption or the release of water molecules from the membrane, confirming the absence of changes in the wetting properties of GO due to electric current across the membrane (dewetting could have caused a decrease in the weight).

We then decreased the humidity of the chamber to 35% RH while an electric current of 20 mA was set across the membrane. In situ weight measurements in this state showed only a small decrease in weight intake (Extended Data Fig. 8b),

indicating retention of the majority of absorbed water molecules inside the GO membrane. This small decrease in weight intake could be attributed to the evaporation of surface-adsorbed water molecules. However, switching the electric current to zero led to continuous release of the retained water molecules from the GO membrane and reached to the initial water intake of the GO membrane at 35% RH. The retention of water molecules inside the GO membrane in the presence of electric current clearly suggests that the current through the membrane limits the diffusion of water molecules inside the GO capillaries rather than changing its wetting properties.

In situ infrared measurements. To monitor the chemical changes in GO membranes during voltage cycling, we performed *in situ* infrared absorption spectroscopy measurements²⁴ in transmission geometry (typically 512 scans per loop) by using VERTEX 80, Bruker FT-IR spectrometer and HYPERION Microscope, using a MCT (mercury cadmium telluride) liquid-N₂-cooled detector with a mirror optical velocity of 0.6329 cm s⁻¹ at a resolution of 4 cm⁻¹. To enable the application of electric potential across the membrane during infrared measurements, we deposited 10-nm Au electrodes on both sides of the freestanding GO membrane. During infrared measurements, the whole system was continuously purged with a dry N₂ stream to remove water in the atmosphere. To match the infrared experimental conditions to those of the permeation experiments, we used a water reservoir (a drop of water) at the edge of the GO membranes (away from the infrared spot) to hydrate the membrane during measurements. Without this reservoir, infrared spectra from the samples resemble that of a dry GO membrane (lower OH vibration peak).

In situ XRD measurements. To probe the changes in the interlayer distance of GO membranes as a function of applied voltage, we performed *in situ* XRD experiments using Bruker D-8 Discover advanced XRD system (Cu K α , $\lambda = 0.154$ nm). A custom-made XRD sample holder was designed to hold a few millilitres of water beneath the GO membrane, providing a continuous source of moisture to keep the membrane at 100% RH, mimicking experimental conditions similar to electrically controlled water permeation. The interlayer spacing d was calculated using Bragg's equation, $d = \lambda/(2\sin\theta)$, where θ is the scattering angle and λ is wavelength of the incident wave.

Electric field due to a current-carrying conductor. To explain electrically controlled water permeation through GO membranes, we propose a simple model. We consider the case of a single conductive filament of length L and radius a ($a \ll L$) carrying a constant stationary current I in a closed circuit with an applied potential V . A current-carrying conductor is known to produce an electric field $E(r, z)$ associated with the electric potential $\psi(r, z)$ around it, depending on its dimensions and conductivity σ ^{27,28,43}. We envisage that this potential $\psi(r, z)$ subsequently decays to zero at a point (at distance b) far from the filament. For any point at a distance r between a and b , Laplace's equation is valid⁴⁴:

$$\nabla^2 \psi(r, z) = 0 \quad (1)$$

The boundary conditions for the above scenario are $\psi(r, z) = -Jz/\sigma$ and $\partial\psi/\partial z = -E_0 = -J/\sigma$ when $r \leq a$ and $\psi(r, z) = 0$ when $r = b$. Here, J is current density through the filament, σ is the conductivity of the wire and z varies from 0 to L (Extended Data Fig. 9a).

Solving equation (1) using these boundary conditions²⁸ in cylindrical coordinates (r, ϕ, z) , we obtain

$$\psi(r, z) = \frac{Jz \ln(r/b)}{\sigma \ln(b/a)} \quad (2)$$

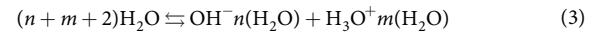
The electric field associated with $\psi(r, z)$ is calculated as

$$E(r, z) = -\left(\frac{\partial\psi}{\partial r} \hat{r} + \frac{\partial\psi}{\partial z} \hat{z}\right) = \frac{Jz}{\sigma r \ln(a/b)} \hat{r} + \frac{J \ln(r/b)}{\sigma \ln(a/b)} \hat{z} \quad (2)$$

At very close distances from the filament, the magnitude of the radial component of the electric field is larger than that of the z component. In our experiments, about 1 nA flows through a single conductive filament of $L \approx 1 \mu\text{m}$ for an applied potential of 1 V. All these filaments are surrounded by the dielectric GO regions, as confirmed from the Raman mapping and AFM experiments. Hence, the electric potential $\psi(r, z)$ decays to zero at some arbitrary radial distance b from the filament. Considering $b = 500$ nm (because the average separation between conductive filaments is about 1 μm , as determined by the Raman measurements) and assuming $a \approx 10$ nm (from PF TUNA imaging), we plot the magnitude of the electric field and its spatial distribution as a function of r and z (Extended Data Fig. 9b). We found that the field remains high close to the filament (less than 10 nm away) and it decays slightly (by about 6 times) from the top positive electrode ($2.3 \times 10^7 \text{ V m}^{-1}$) to a point a distance 100 nm above the bottom electrode ($4 \times 10^6 \text{ V m}^{-1}$). The field around the filament persists even up to a radius of 50 nm, with a magnitude varying from $4.3 \times 10^6 \text{ V m}^{-1}$ to $5.4 \times 10^5 \text{ V m}^{-1}$ depending on the distance from

the top positive electrode. We also found that for the same 1-nA current, the purely radial component of E decreases by about 40 times as the radius of the filament is increased from 10 nm to 100 nm. In summary, the current-carrying filaments in GO membranes produce a radial electric field that is sufficiently strong to dissociate water molecules into OH⁻ and H₃O⁺ ions. Importantly, all of the above estimates are based on a simple model of a single straight conducting wire (zero-order approximation). However, the complicated structure of conducting filaments could produce even higher electric fields owing to the close arrangement of individual filaments, especially near the positive electrode.

Electric-field-induced dissociation of water. Electric-field-associated dissociation of water^{13,45-47} (without producing O₂ and H₂ gas) into H₃O⁺ and OH⁻ ions has been observed previously, even in the presence of transitory fields generated by molecular fluctuations⁴⁵. On the basis of this, we propose that the strong electric field near the conducting filaments in the GO membranes dissociates water in the interlayer channels into OH⁻ and H₃O⁺ ions according to



By increasing the electric current in conducting filaments, the electric field (equation (2)) increases and hence the water dissociation rate increases. Once the electric current is switched off, the field vanishes and the reaction kinetics towards the left-hand side of equation (3) becomes more favourable.

The rate of reaction for the ionization in equation (3) depends on the activation energy. This energy value is known to decrease in the presence of an electric field⁴⁷ owing to the induced local dipoles in the system, and the water dissociation rate in the presence of electric field is

$$k_D(E) = A \exp\left[-\frac{Q(E)}{k_B T}\right]$$

where A is pre-exponential factor, $Q(E)$ is field-dependent activation energy for dissociation, k_B is the Boltzmann constant and T is temperature. In a simple form, $Q(E)$ can be expressed as $Q(0) - \Delta Q$, where $Q(0)$ is the field-free activation energy for dissociation and ΔQ is the field-induced decrease in activation energy. Thus, k_D increases with increasing electric field. The permeation experiments show that the water permeation rate P decreases as the current-induced electric field around the conducting filaments is increased. Therefore, $P \propto 1/k_D$ and $P(0)/P(E) = \exp[\Delta Q/(k_B T)]$. From the observed change in P due to the electrical current flow (for example, 20 mA) in our experiments, we estimated the decrease in activation energy for dissociation of water molecules and to be nearly 12% with respect to the activation energy at room temperature in the absence of an electric field.

Molecular dynamics simulations. To understand the influence of the dissociation of water on the water flow through the GO membranes, we performed non-equilibrium molecular dynamics simulations. Water permeation through GO membranes is believed to occur along a network of pristine graphene channels that develop between functionalized areas of GO sheets¹⁷ (typically, an area of 40–60% remains free from functionalization^{48,49}). Thus, we performed molecular dynamics simulations using a large-scale atomic/molecular massively parallel simulator (LAMMPS)⁵⁰, to investigate the dynamical properties (flow rate) of a mixture of H₃O⁺, OH⁻ and water inside pristine graphene capillaries¹⁷ (Extended Data Fig. 10a).

For molecular dynamics simulations, a rigid model was used for H₃O⁺ and OH⁻ ions, as presented in previous studies^{51,52}. The graphene layers were kept fixed and an SPC/E model was used to describe the water molecules. The carbon and oxygen atoms interact via Lennard-Jones pair potentials ($\sigma_C = 0.0553 \text{ kcal mol}^{-1}$, $\varepsilon_C = 3.4 \text{ \AA}$; $\sigma_{\text{H}_3\text{O}^+} = 0.147467 \text{ kcal mol}^{-1}$, $\varepsilon_{\text{H}_3\text{O}^+} = 3.05 \text{ \AA}$; $\sigma_{\text{OH}^-} = 0.149618 \text{ kcal mol}^{-1}$, $\varepsilon_{\text{OH}^-} = 3.84 \text{ \AA}$), and cross-Lennard-Jones potential parameters were obtained from the Lorentz–Berthelot combining rules. The cut-off radius for the Lennard-Jones potential was chosen as 10 \AA . The *NVT* ensemble (Nosé–Hoover thermostat) was used to control the temperature for both non-rigid and rigid molecules at room temperature. A particle–particle particle–mesh was used to compute the long-range Coulomb interaction with a desired relative error in the forces for long-range Coulomb interaction solvers of 10^{-4} . In all cases, the time step was chosen as 1 fs. The force fields were validated by calculating the diffusion coefficient of bulk water, $D_0 = 2.45 \times 10^{-5} \text{ cm}^2 \text{ s}^{-1}$, which is in good agreement with previous experimental results⁵³.

Extended Data Fig. 10a shows the simulation box, which contains two compartments that are connected by a graphene capillary of height 10 \AA . Periodic boundary conditions were applied along the z direction in the boxes and along the y direction in the capillary. The simulation unit cell contained 4,362 water or ion molecules and each graphene capillary has a size of 8 nm \times 2 nm (720 carbon atoms). First, the molecules inside the left box (Extended Data Fig. 10a) were relaxed to enable it to reach its equilibrium for 1 ns. Afterwards, we applied a pressure of 0.9 bar on the vertical wall of the left box (shown by the arrow in Extended Data Fig. 10a) to move the wall towards the capillary, so that the water and ion molecules enter

the capillary and flow towards the right box. The rate of filling of the capillary and the right box was found to depend strongly on the number of H_3O^+ and OH^- ions in the system (Extended Data Fig. 10b, c). We further found that, after the equilibration, the concentration of ions inside the capillary was slightly lower than the concentration of ions in the system (left box), owing to the higher water flow to the capillary. We calculated the concentration of ions inside the capillary with respect to the number of water molecules in it for different total concentrations of ions in the system and studied the influence of this on the water flow rate through the capillary. We found that by increasing the concentration of ions inside capillary, the water flow rate through the capillary decreased. When the ion concentration reached more than 6% (by number), the water flow rate was reduced substantially (about 30 times; Extended Data Fig. 10d). This is qualitatively in agreement with the experimental estimate of the ion concentration (see Methods section 'Influence of H_3O^+ and OH^- ions'); however, the discrepancy in the exact values is expected because realistic GO channels contain functionalities, rough edges and so on, which are difficult to model accurately. Also, the concentration of ions calculated in molecular dynamics simulations for substantial reduction of water flow rate could be an overestimate. The free space available in the interlayer channels of GO is less than 1 nm, the size used in molecular dynamics simulations, and the functional groups inside GO capillaries may absorb ions, making the channel more hydrophilic. These effects were not included in the molecular dynamics simulations. Nevertheless, our molecular dynamics simulations suggest that the dissociated water molecules inside the interlayer channels of GO membranes could substantially affect water permeation through the membrane.

The observed decrease in water flow rate with increasing ion concentration in the capillary could be due to ion hydration effects: with an increasing number of ions inside graphene capillary, water tends to remain inside the capillary thereby decreasing the water flow. A high concentration of dissociated water (about 50%) was also observed at the interfaces and metallic surfaces, demonstrating that they could be energetically stable even at room temperature⁵⁴.

Interaction of H_3O^+ and OH^- ions with water and graphene or graphene oxide capillaries. First-principles calculations were performed using Gaussian software and the B3LYP/6-31g(d) level of theory to understand the interaction of H_3O^+ and OH^- ions with GO or graphene surfaces and the surrounding water molecules. First, we optimized two GO flakes: each contains 62 carbon atoms, 22 hydrogen atoms, two epoxy and three hydroxyl groups. Then, we added 28 water molecules between them and re-optimized the system. The final configuration at this stage showed the formation of a few hydrogen bonds between confined water molecules and epoxy or hydroxyl functional groups. Later, we replaced two water molecules located at the centre of two GO sheets by one H_3O^+ and one OH^- ion and re-optimized the system. It is evident from the resulted configuration (Fig. 4a) that H_3O^+ and OH^- ions form big clusters with the surrounding water molecules, owing to substantial changes in the hydrogen-bond network. We also performed similar first-principles calculations for H_3O^+ and OH^- ions in between two pristine graphene sheets and found large hydrated cluster formation with surrounding water molecules (Fig. 4b). This additional calculation confirms that the chemistry of the capillary surface has no substantial influence on the interaction of ions with water.

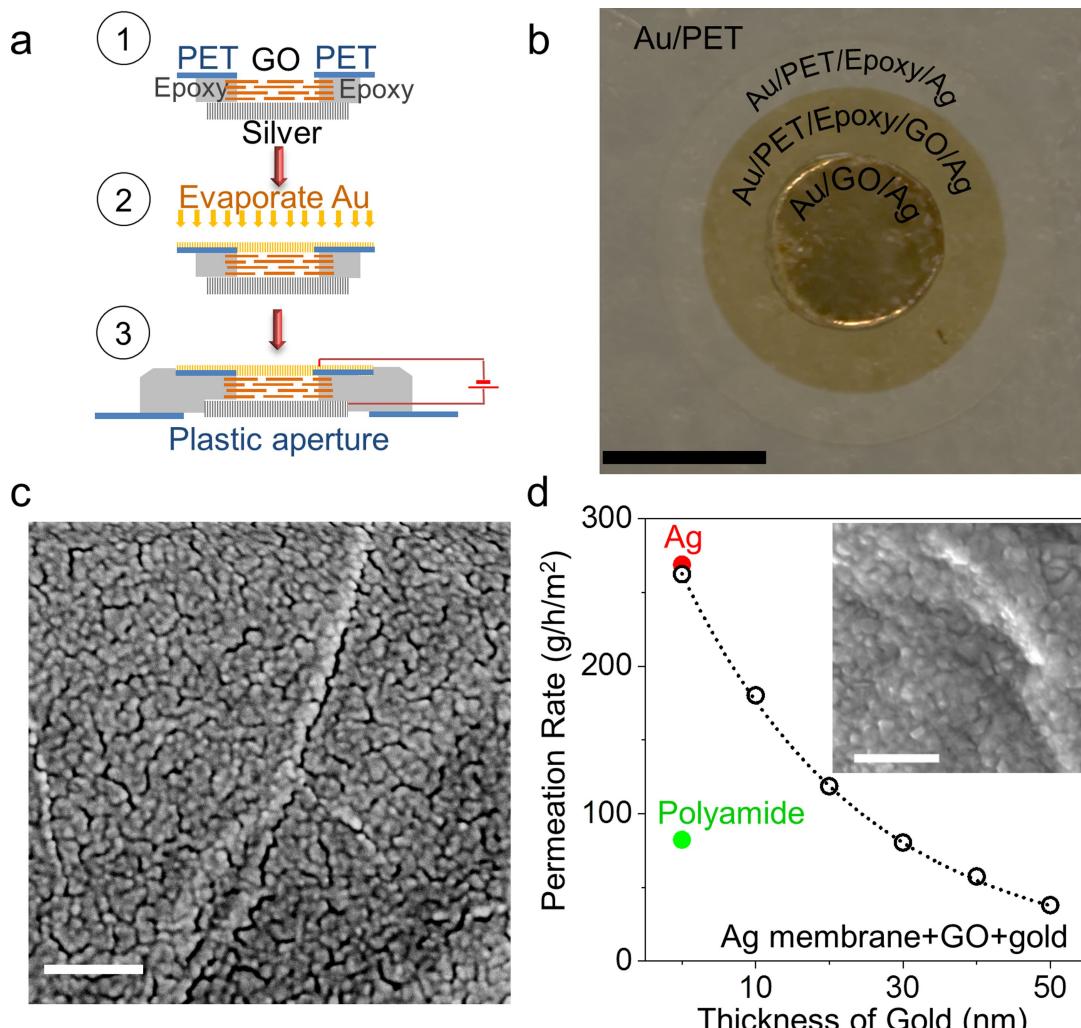
The first hydration shell of both H_3O^+ and OH^- ions (Fig. 4c, d) is found to be unique and form a highly ordered hydrogen-bonded water structure, in agreement with previous modelling⁵⁵. In addition, ions form hydrogen bonds with epoxy or hydroxyl groups if they are close to the GO surface. However, this phenomenon is found to be less probable owing to the presence of water molecules around the H_3O^+ and OH^- ions. Our first-principles modelling indicates that the experimentally observed reduction in water permeation in the presence of H_3O^+ and OH^- ions is due to their interaction with surrounding water molecules (via strong hydrogen bonds), which creates large hydrated clusters, blocking the water transport. It has been reported previously that the diffusion coefficient of water decreases substantially in these large clusters of hydrated H_3O^+ ions⁵⁵. In addition, the reaction dynamics of the creation or recombination of H_3O^+ and OH^- ions (equation (3)) in the presence of an electric field could further contribute to the observed decrease in water permeation, owing to the perturbations caused in the hydrogen-bonded network of water inside the graphene capillary¹⁷. Any disorder in the ordered hydrogen-bonded structure of water is also expected to affect the slip-enhanced flow of water through the graphene capillary.

Influence of H_3O^+ and OH^- ions. To study the influence of H_3O^+ and OH^- ions on the water permeation rate, we carried out additional osmotic-pressure-driven water permeation experiments using GO membranes. These experiments were carried out using a similar experimental set-up to that shown in Extended Data Fig. 7a, under same experimental conditions, but adding NaOH and HCl solutions of the same concentration (to maintain the same osmotic pressure) in both the water and sucrose compartments. Here, the changes in water permeation rate were monitored with respect to pure water flux when the acid and base solution was absent. To eliminate complications in the following discussion, we denote the

solution added in the water (sucrose) compartments as S_w (S_s). Figure 4e shows the water permeation rate for S_s and S_w of 10 mM concentration. When S_s and S_w are both 10 mM NaOH the water flux is increased to approximately $0.15 \text{ l h}^{-1} \text{ m}^{-2}$ with respect to a pure water flux of $0.1 \text{ l h}^{-1} \text{ m}^{-2}$. However, it is decreased to approximately $0.075 \text{ l h}^{-1} \text{ m}^{-2}$ when S_s and S_w are 10 mM HCl. The small decrease (increase) in the water flux in the case of HCl (NaOH) could be attributed to the pH-dependant decrease (increase) in the interlayer spacing of GO membranes⁵⁶. However, the water flux is reduced substantially to approximately $0.01 \text{ l h}^{-1} \text{ m}^{-2}$ when using 10 mM NaOH and 10 mM HCl as S_s and S_w , respectively. Figure 4f further shows the decrease in the water permeation rate as the concentration increases from 1 mM to 10 mM, when S_s and S_w are NaOH and HCl, respectively. These experiments further confirm the demand for the simultaneous presence of both H_3O^+ and OH^- ions in controlling the water permeation in GO membranes. Interestingly, the water permeation rate recovered to the original value when the acid and alkaline solutions were substituted with pure water and sucrose solutions. As a control experiment, we also performed measurements with one compartment filled with 10 mM NaOH or HCl while keeping the other compartment free from acid or base. No substantial reduction in the water permeation rate was observed in this case. Similarly, we performed experiments with $S_s = S_w = 10 \text{ mM NaCl}$ and did not find any noticeable change in the water permeation rate.

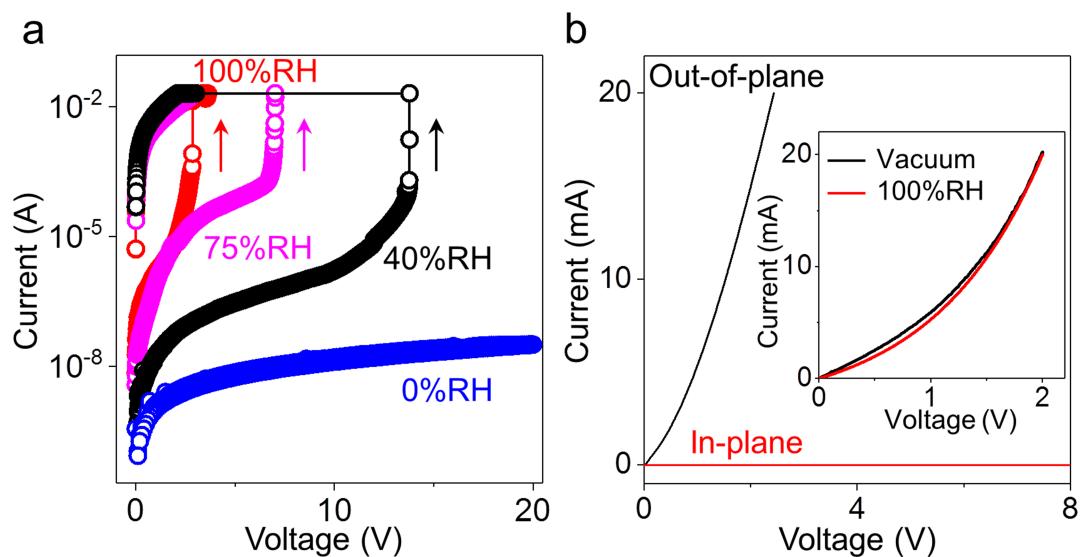
Data availability. The data that support the findings of this study are available from the corresponding authors on reasonable request. Data related to molecular dynamics simulations are available from M.N.-A. (mehdi.neekamal@gmail.com).

31. Siegel, J., Lyutakov, O., Rybka, V., Kolská, Z. & Svorčík, V. Properties of gold nanostructures sputtered on glass. *Nanoscale Res. Lett.* **6**, 96 (2011).
32. O'Dwyer, J. J. Dielectric breakdown in solids. *Adv. Phys.* **7**, 349–394 (1958).
33. Kim, S. K. et al. Conductive graphitic channel in graphene oxide-based memristive devices. *Adv. Funct. Mater.* **26**, 7406–7414 (2016).
34. Eda, G. et al. Graphene oxide gate dielectric for graphene-based monolithic field effect transistors. *Appl. Phys. Lett.* **102**, 133108 (2013).
35. Standley, B., Mendez, A., Schmidgall, E. & Bockrath, M. Graphene-graphite oxide field-effect transistors. *Nano Lett.* **12**, 1165–1169 (2012).
36. Lee, J. S., Lee, S. & Noh, T. W. Resistive switching phenomena: a review of statistical physics approaches. *Appl. Phys. Rev.* **2**, 031303 (2015).
37. Qin, S. et al. A physics/circuit-based switching model for carbon-based resistive memory with sp^2/sp^3 cluster conversion. *Nanoscale* **4**, 6658–6663 (2012).
38. Chen, C. et al. Annealing a graphene oxide film to produce a free standing high conductive graphene film. *Carbon* **50**, 659–667 (2012).
39. Borini, S. et al. Ultrafast graphene oxide humidity sensors. *ACS Nano* **7**, 11166–11173 (2013).
40. Pei, S. & Cheng, H. The reduction of graphene oxide. *Carbon* **50**, 3210–3228 (2012).
41. Park, S. et al. Colloidal suspensions of highly reduced graphene oxide in a wide variety of organic solvents. *Nano Lett.* **9**, 1593–1597 (2009).
42. Ganguly, A., Sharma, S., Papakonstantinou, P. & Hamilton, J. Probing the thermal deoxygenation of graphene oxide using high-resolution *in situ* X-ray-based spectroscopies. *J. Phys. Chem. C* **115**, 17009–17019 (2011).
43. Müller, R. A semiquantitative treatment of surface charges in DC circuits. *Am. J. Phys.* **80**, 782–788 (2012).
44. Jackson, J. D. *Classical Electrodynamics* 3rd edn, Ch. 1, 12–14 (John Wiley & Sons, New York, 1999).
45. Geissler, P. L., Dellago, C., Chandler, D., Hutter, J. & Parrinello, M. Autoionization in liquid water. *Science* **291**, 2121–2124 (2001).
46. Mafé, S., Ramírez, P. & Alcaraz, A. Electric field-assisted proton transfer and water dissociation at the junction of a fixed-charge bipolar membrane. *Chem. Phys. Lett.* **294**, 406–412 (1998).
47. Pinkerton, T. D. et al. Electric field effects in ionization of water–ice layers on platinum. *Langmuir* **15**, 851–856 (1999).
48. Wilson, N. R. et al. Graphene oxide: structural analysis and application as a highly transparent support for electron microscopy. *ACS Nano* **3**, 2547–2556 (2009).
49. Loh, K. P., Bao, Q., Eda, G. & Chhowalla, M. Graphene oxide as a chemically tunable platform for optical applications. *Nat. Chem.* **2**, 1015–1024 (2010).
50. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
51. Vácha, R., Buch, V., Milet, A., Devlin, J. P. & Jungwirth, P. Autoionization at the surface of neat water: is the top layer pH neutral, basic, or acidic? *Phys. Chem. Chem. Phys.* **9**, 4736–4747 (2007).
52. Vácha, R., Horinek, D., Berkowitz, M. L. & Jungwirth, P. Hydronium and hydroxide at the interface between water and hydrophobic media. *Phys. Chem. Chem. Phys.* **10**, 4975–4980 (2008).
53. Mills, R. Self-diffusion in normal and heavy water in the range 1–45.deg. *J. Phys. Chem.* **77**, 685–688 (1973).
54. Meyer, B. et al. Partial dissociation of water leads to stable superstructures on the surface of zinc oxide. *Angew. Chem. Int. Ed.* **43**, 6641–6645 (2004).
55. Brodskaya, E., Alexander, P. L. & Aatto, L. Investigation of water clusters containing OH^- and H_3O^+ ions in atmospheric conditions. A molecular dynamics simulation study. *J. Phys. Chem. B* **106**, 6479–6487 (2002).
56. Huang, H. et al. Salt concentration, pH and pressure controlled separation of small molecules through lamellar graphene oxide membranes. *Chem. Commun.* **49**, 5963–5965 (2013).


Extended Data Fig. 1 | Metal-Go-metal sandwiched membranes.

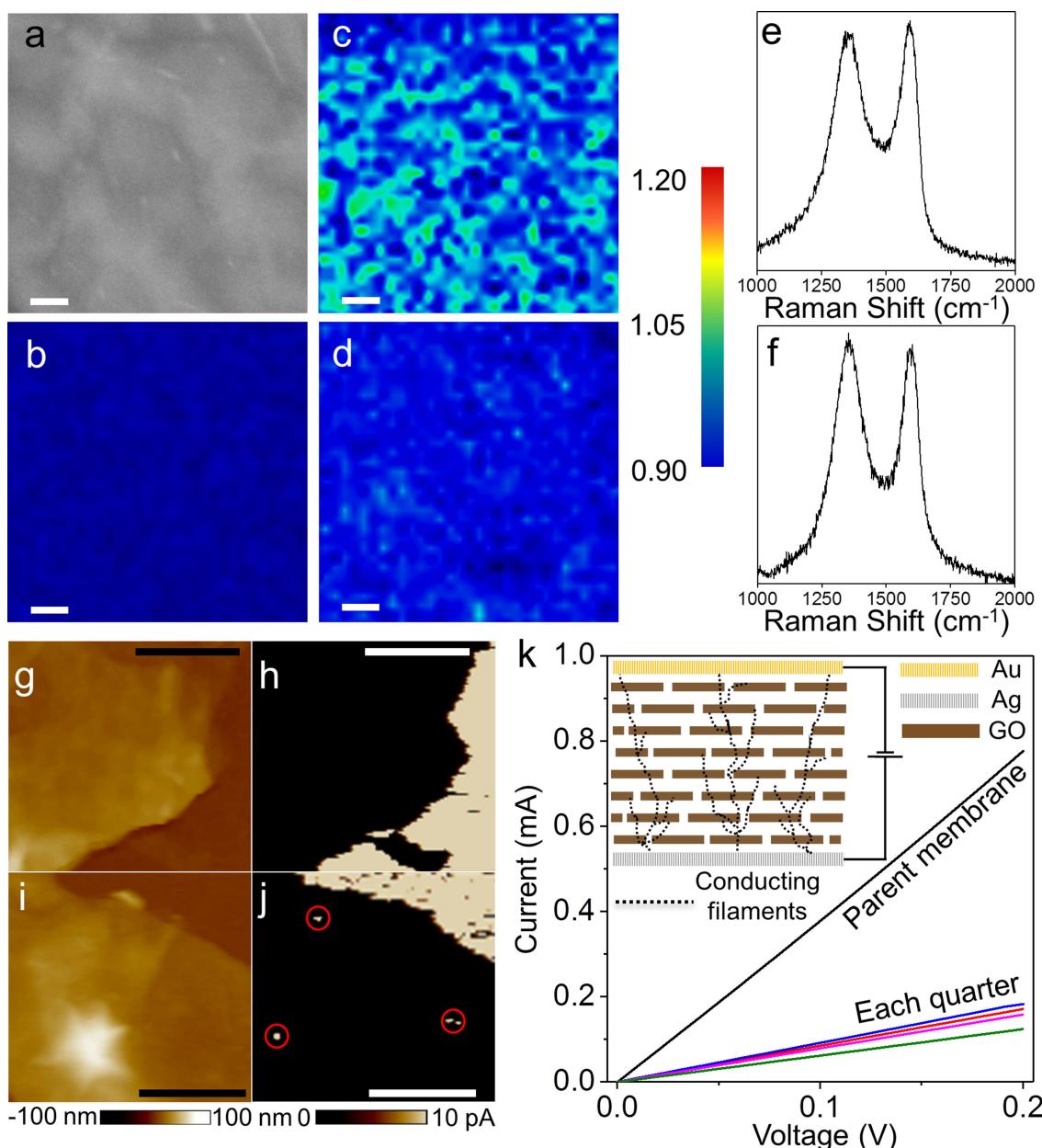
a, Fabrication procedure for the metal-Go-metal sandwich membrane. **b**, Photograph of one of our metal-Go-metal sandwich membranes attached to the PET sheet (step 2 in **a**). Scale bar, 6 mm. This was further attached onto another plastic disk to seal the metal container for gravimetric testing. **c**, SEM image showing the discontinuities and voids in a 10-nm gold thin film on a GO membrane. Scale bar, 150 nm. **d**, Water

permeation rate of metal-Go-metal sandwiched membranes as a function of gold electrode thickness. The dotted line is a guide to the eye. Water permeation rates of a bare porous silver (Ag) support (red filled circle) and a commercial polyamide nanofiltration membrane (green filled circle) are provided for comparison. Inset, SEM image of a 50-nm-thick gold thin film on a GO membrane. Scale bar, 150 nm.



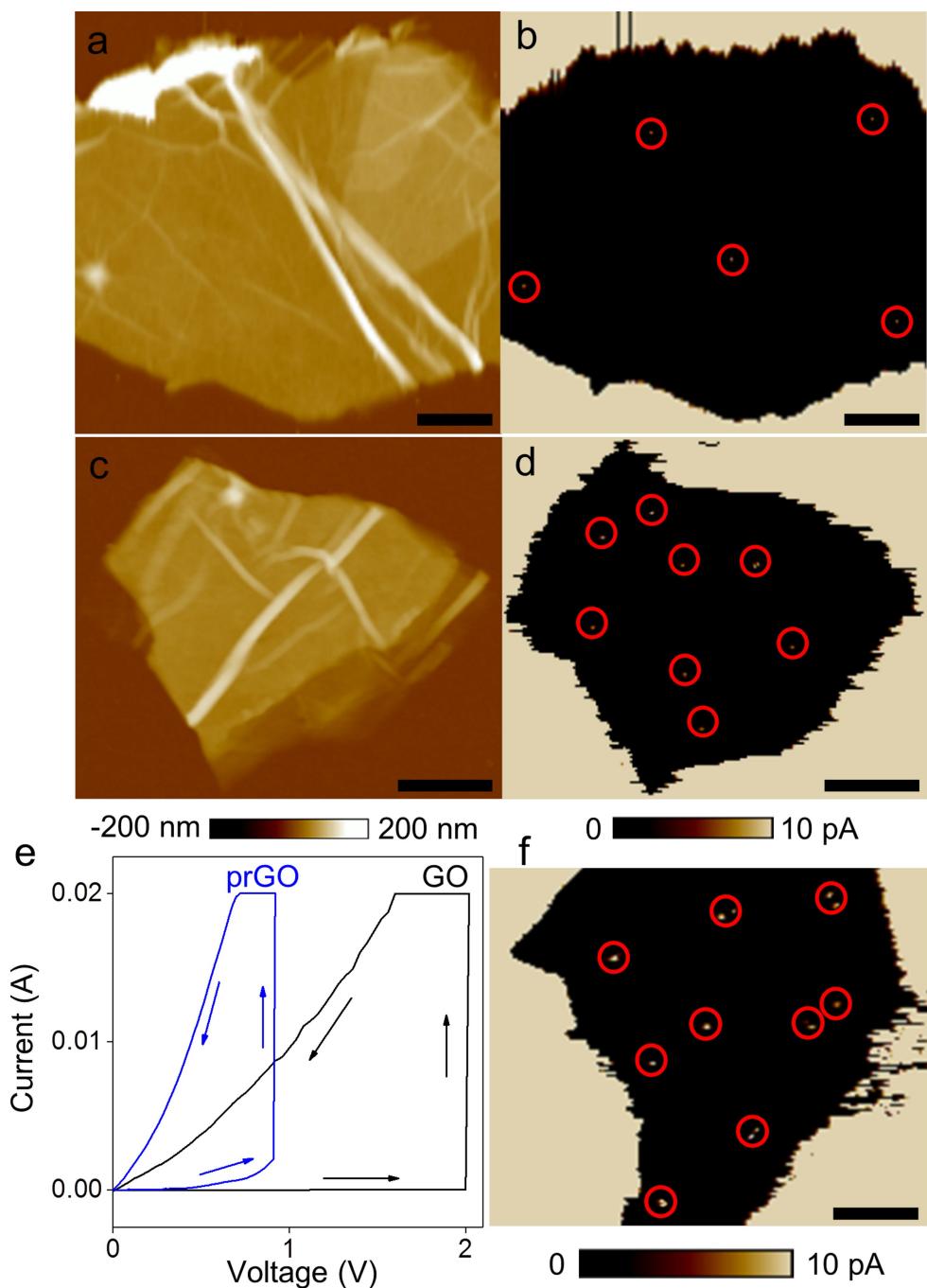
Extended Data Fig. 2 | Conducting filament formation in a GO membrane and its electrical characterization. **a**, I – V characteristics during the first voltage sweep show a sudden increase in the current for membranes exposed to humid conditions, suggesting partial electrical

breakdown of the GO membrane and conducting filament formation. **b**, In-plane and out-of-plane I – V characteristics of the GO membrane after filament formation. Inset, out-of-plane I – V characteristics of the GO membrane at 100% RH and vacuum.



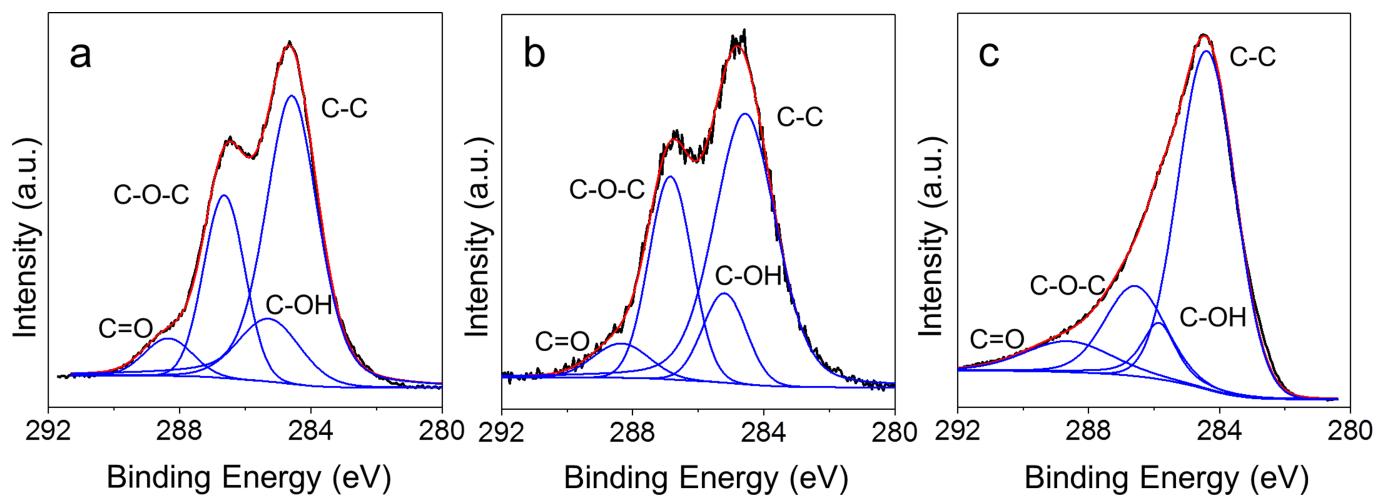
Extended Data Fig. 3 | Raman and AFM characterization of conducting filaments in GO membranes. **a**, Topographical SEM image of a GO membrane after the formation of conducting filaments. **b–d**, Raman intensity ratio (I_D/I_G) mapping of D and G bands for a pristine GO membrane (**b**) and a GO membrane after conducting filaments have formed (**c, d**). **c**, Raman imaging from the membrane surface close to the positive electrode (about 200 nm away). **d**, Raman imaging from the membrane surface close to the negative electrode (about 100 nm away). **e, f**, The Raman spectra from the dark blue and green regions in **c**,

respectively. **g–j**, Topography and the corresponding TUNA current image of pristine GO (**g** and **h**, respectively) and conducting GO (**i** and **j**) membranes (filament formed at 100% RH) exfoliated on a gold-thin-film-coated Si substrate. The conducting filaments are marked by red circles. Scale bars, 1 μm . **k**, Out-of-plane I - V characteristics of a conducting GO membrane with a diameter of about 7 mm before (parent membrane) and after dividing into four equal pieces. Inset, schematic of the structure of conducting carbon filaments in the GO membrane.



Extended Data Fig. 4 | Influence of intercalated water and oxygen content on conducting filament formation in GO membranes.
a, b, Topography and the corresponding TUNA current images of GO membranes after filament formation at 40% RH (**a** and **b**, respectively) and inside liquid water (**c** and **d**). **e**, I – V characteristics of pristine and partially

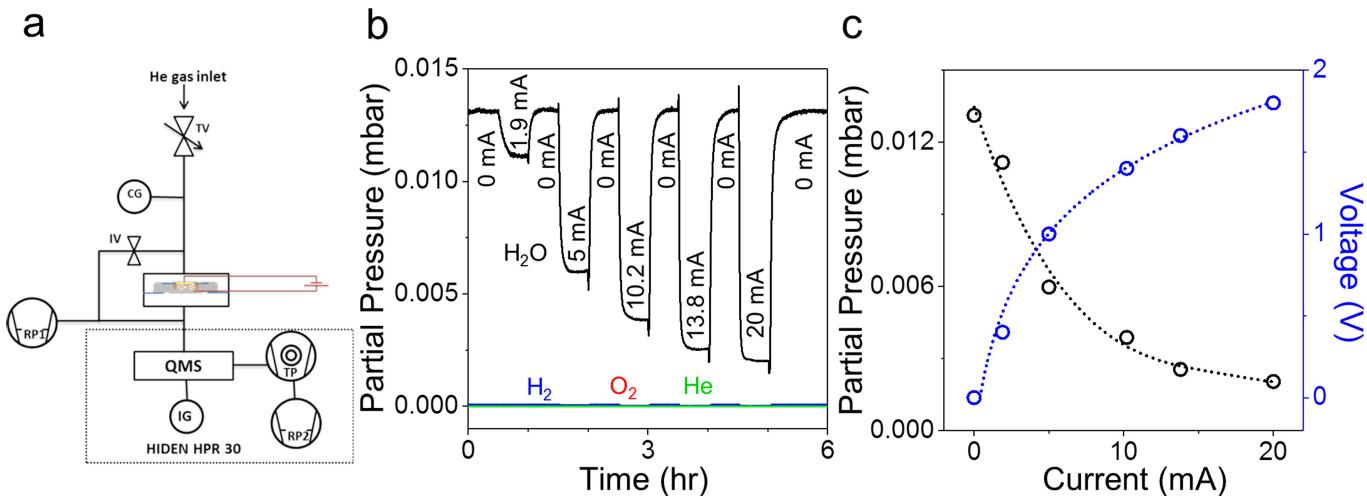
reduced ('pr') GO membranes during the first voltage sweep at 100% RH show partial breakdown. **f**, TUNA current image of a partially reduced GO membrane after filament formation. The conducting filaments are marked by red circles. Scale bars, 2 μ m.



Extended Data Fig. 5 | XPS characterization of GO membranes.

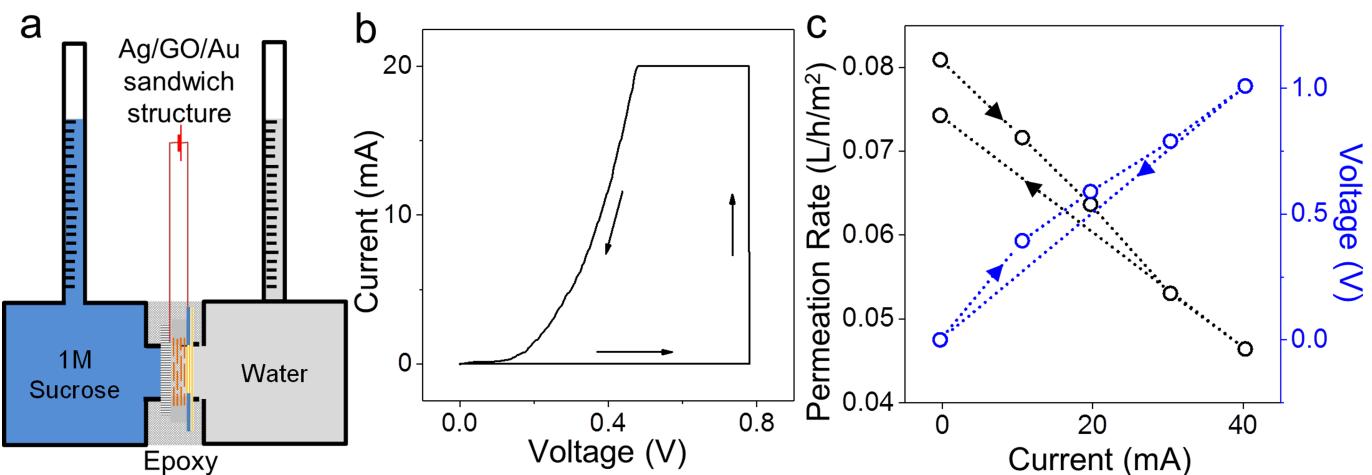
a, C 1s spectrum from a pristine GO membrane. **b, c**, C 1s spectra from GO membranes used for the electrically controlled permeation experiments after filament formation, from a freshly cleaved membrane

surface close to the inner middle region (**b**) and close to the positive electrode (**c**). Black lines, raw data; red lines, the fitting envelope; blue lines, deconvolved peaks attributed to the chemical environments indicated.



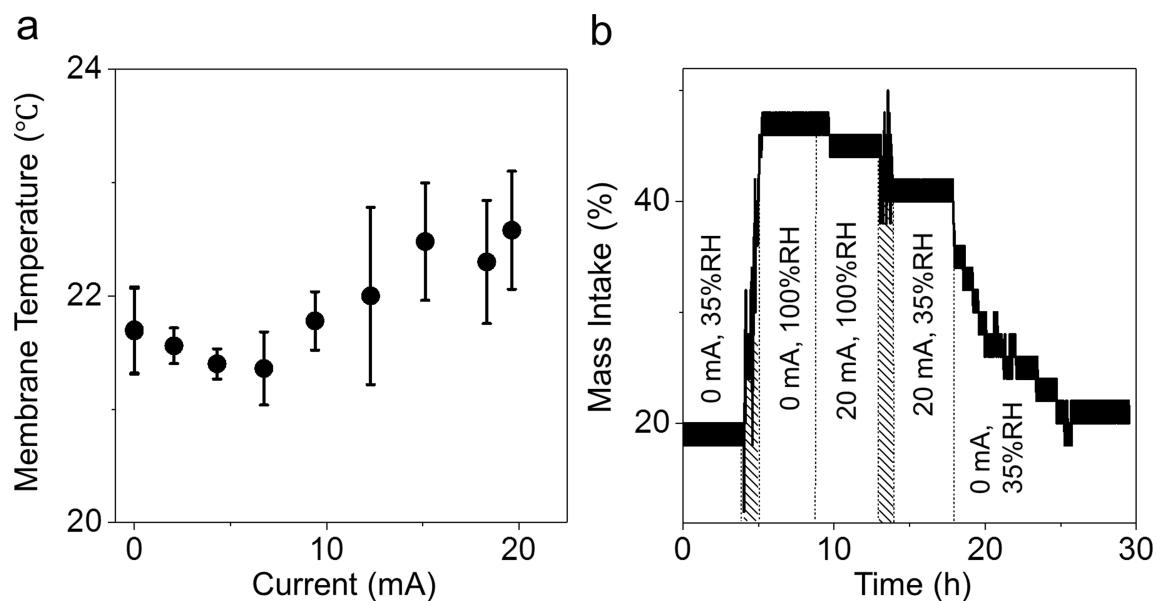
Extended Data Fig. 6 | Mass spectrometry to probe electrically controlled water permeation. **a**, Schematic of the experimental set-up for mass spectrometry measurements. A throttle valve (TV) controls the gas inlet with a capacitance gauge (CG) used to measure the upstream pressure. An isolation valve (IV) isolates upstream and downstream sides of the membrane. A rotary pump (RP1) evacuates the feed and the permeate side to 1 mbar. The quadrupole mass spectrometer (QMS) measures the downstream partial pressure. A turbomolecular pump (TP) backed by a rotary pump (RP2) evacuates the high-vacuum chamber of

the mass spectrometer. An active ion gauge (IG) measures the pressure down to 1×10^{-9} torr in the high-vacuum side. **b**, The partial pressure of He, H₂, O₂ and H₂O at the permeate side as a function of time at different currents through the membrane. No detectable change is observed in the partial pressures of He, H₂ and O₂ under different currents through the membrane. **c**, The partial pressure of H₂O as a function of the current across the GO membrane and the corresponding I-V characteristics (colour-coded axes). The dotted lines are guides to the eye.



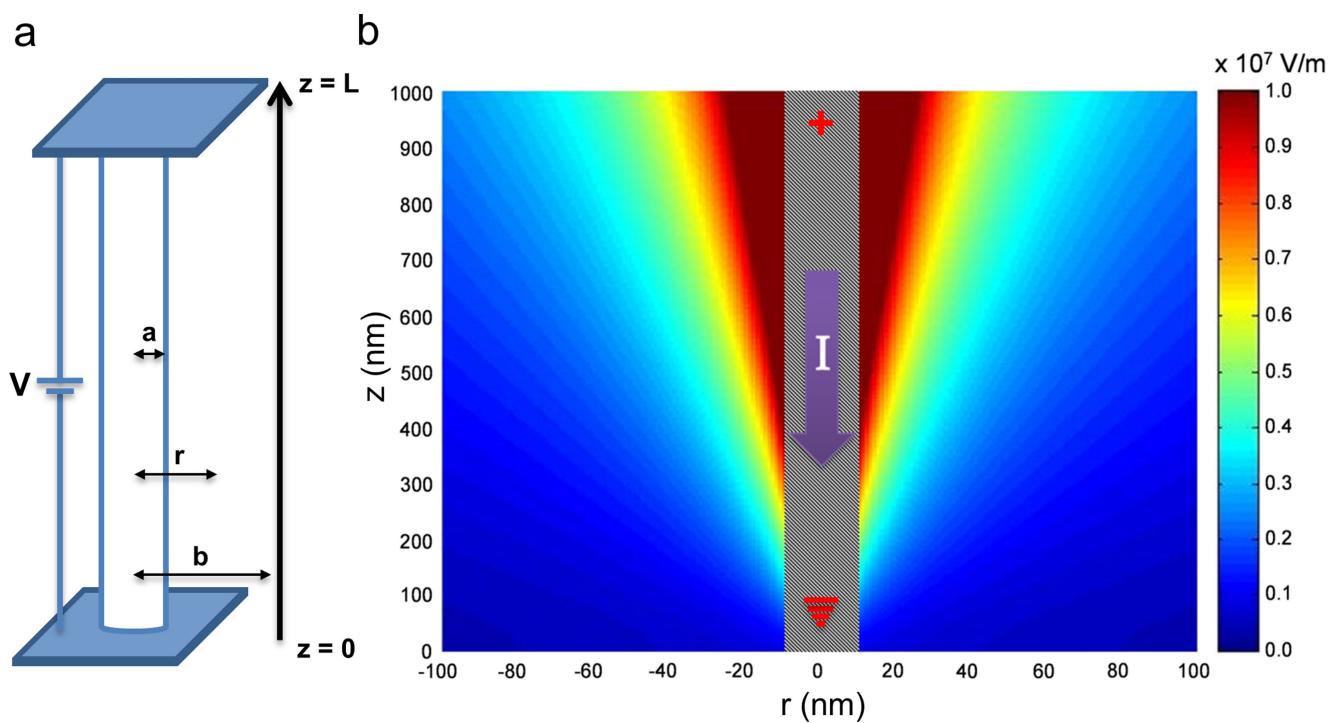
Extended Data Fig. 7 | Electrically controlled liquid water permeation in GO membranes. **a**, Schematic of the experimental set-up. **b**, I – V characteristics of a Au/GO/Au membrane during the first voltage sweep while it is immersed in liquid water in the experimental set-up. **c**, Liquid

water permeation rate as a function of current across the membrane after filament formation and the corresponding I – V characteristics (colour-coded axes). Sample-to-sample variation in the permeation is less than 30% (three samples measured).



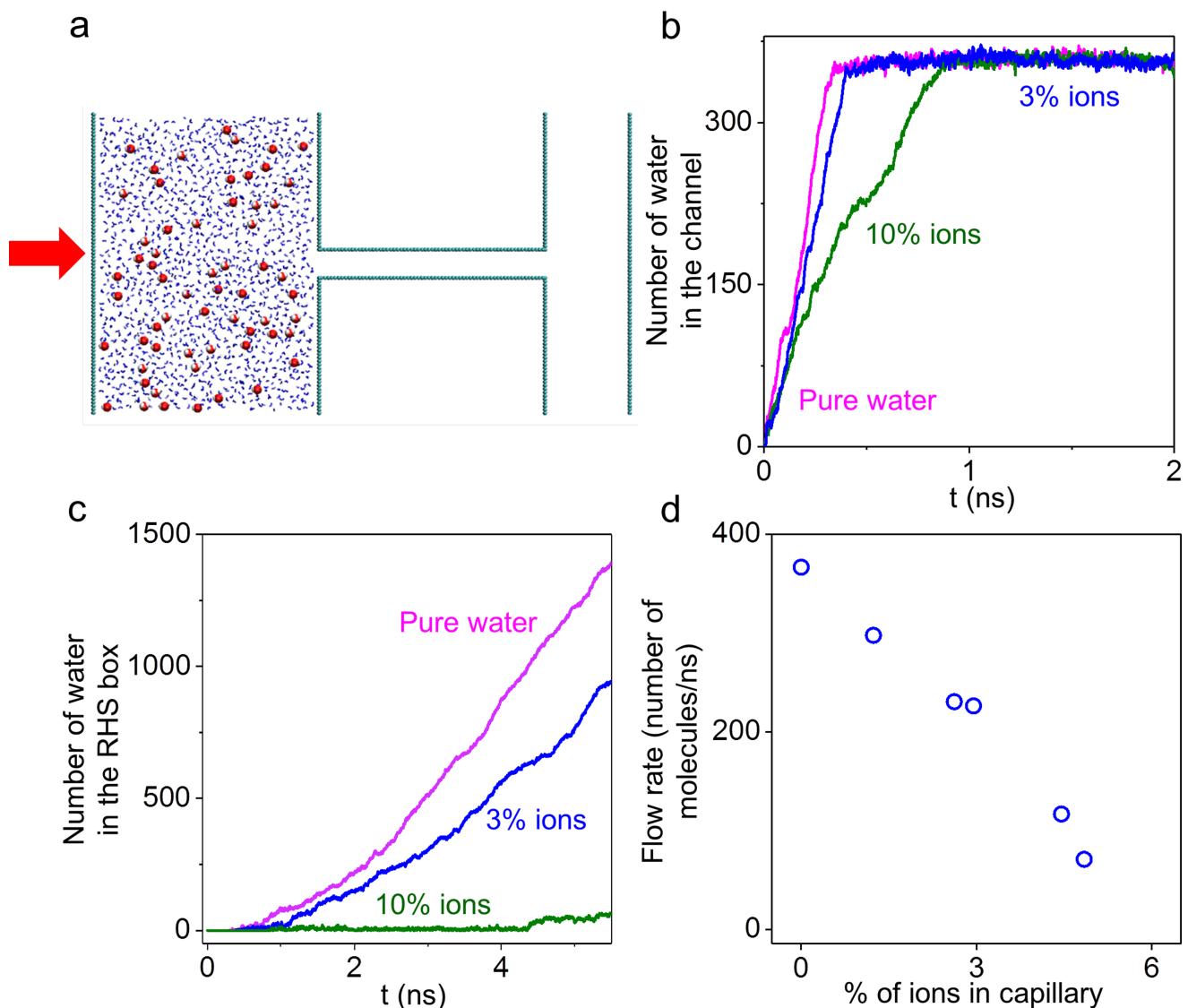
Extended Data Fig. 8 | In situ membrane temperature and water absorption measurements. **a**, Measured membrane temperature as a function of the current flowing across the membrane during the electrically controlled water permeation experiment. Error bars, standard deviation from 10 different measurements across the sample. **b**, The

weight intake of a Au/GO/Ag membrane (1- μm -thick GO) at different humidity and electric current values. Weight intake is calculated with respect to the weight of the membrane at 0% RH. The shaded areas show the time during humidity sweeps.



Extended Data Fig. 9 | Electric field around a current-carrying conductor. **a**, Schematic of the application of a voltage V across an electrically conducting wire with radius a and length L ; b is the point at which the potential decays to zero; r represents any point between a

and b where the electric field E is calculated. **b**, Magnitude of E and its spatial distribution as a function of r and z around a conductive filament with 1-V potential difference across the ends and with 1-nA current flow.



Extended Data Fig. 10 | Molecular dynamics simulations. **a**, Side view of our molecular dynamics simulation set-up used to study the flow of water mixed with H_3O^+ and OH^- ions in the graphene capillary. The model contains two boxes connected by a graphene capillary. At the beginning of the simulation, water was mixed with H_3O^+ and OH^- ions (red and white dots). By moving the left wall (subjected to external pressure) of the box towards the capillary, the water flow is created and the right box is

gradually filled. The arrow indicates the direction of the external pressure applied on the left wall of the box. **b, c**, Number of water molecules in the capillary (**b**) and number of water molecules in the right box (**c**) for pure water and water with ions once pressure is applied to the left box (colour-coded labels). **d**, Water flow rate as a function of the concentration of ions inside the capillary.

North Pacific freshwater events linked to changes in glacial ocean circulation

E. Maier^{1*}, X. Zhang^{1*}, A. Abelmann¹, R. Gersonde¹, S. Mulitza², M. Werner¹, M. Méheust¹, J. Ren¹, B. Chaplgin³, H. Meyer³, R. Stein¹, R. Tiedemann¹ & G. Lohmann¹

There is compelling evidence that episodic deposition of large volumes of freshwater into the oceans strongly influenced global ocean circulation and climate variability during glacial periods^{1,2}. In the North Atlantic region, episodes of massive freshwater discharge to the North Atlantic Ocean were related to distinct cold periods known as Heinrich Stadials^{1–3}. By contrast, the freshwater history of the North Pacific region remains unclear, giving rise to persistent debates about the existence and possible magnitude of climate links between the North Pacific and North Atlantic oceans during Heinrich Stadials^{4,5}. Here we find that there was a strong connection between changes in North Atlantic circulation during Heinrich Stadials and injections of freshwater from the North American Cordilleran Ice Sheet to the northeastern North Pacific. Our record of diatom $\delta^{18}\text{O}$ (a measure of the ratio of the stable oxygen isotopes ^{18}O and ^{16}O) over the past 50,000 years shows a decrease in surface seawater $\delta^{18}\text{O}$ of two to three per thousand, corresponding to a decline in salinity of roughly two to four practical salinity units. This coincided with enhanced deposition of ice-rafted debris and a slight cooling of the sea surface in the northeastern North Pacific during Heinrich Stadials 1 and 4, but not during Heinrich Stadial 3. Furthermore, results from our isotope-enabled model⁶ suggest that warming of the eastern Equatorial Pacific during Heinrich Stadials was crucial for transmitting the North Atlantic signal to the northeastern North Pacific, where the associated subsurface warming resulted in a discernible freshwater discharge from the Cordilleran Ice Sheet during Heinrich Stadials 1 and 4. However, enhanced background cooling across the northern high latitudes during Heinrich Stadial 3—the coldest period in the past 50,000 years⁷—prevented subsurface warming of the northeastern North Pacific and thus increased freshwater discharge from the Cordilleran Ice Sheet. In combination, our results show that nonlinear ocean-atmosphere background interactions played a complex role in the dynamics linking the freshwater discharge responses of the North Atlantic and North Pacific during glacial periods.

During the last glacial period (roughly 115,000 to 12,000 years ago), large parts of the North American continent were covered by the North American Ice Sheet Complex, which comprised the Laurentide Ice Sheet (LIS) in the centre and east, and the smaller Cordilleran Ice Sheet (CIS) in the west (Fig. 1). The LIS was a source of major freshwater discharge events into the North Atlantic Ocean during Heinrich Stadials. It has been proposed that the resulting weakening of the Atlantic meridional overturning circulation (AMOC)^{3,8} also influenced circulation in the North Pacific Ocean^{5,9}. However, the dynamical connections between glacial ocean circulation changes and CIS dynamics remain elusive.

Because freshwater is a major modulator of ocean stratification and hence vertical mixing, reconstructions of North Pacific freshwater flux history help in elucidating the evolution of climate in the North Pacific. During the last glacial, the ice volume of the CIS was around 4–16 times larger than it is today¹⁰, making it a major source of freshwater for the northeastern North Pacific. To date, however, palaeoclimate

reconstructions have shown conflicting results with respect to freshwater flux from the CIS. Large deposits of ice-rafted debris (IRD) have been observed in a few coastal settings of the northeastern North Pacific during Heinrich Stadials, indicating episodic freshwater input from melting icebergs¹¹. However, $\delta^{18}\text{O}$ records from planktic foraminifera ($\delta^{18}\text{O}_{\text{pl.foram.}}$) in the open northeastern North Pacific are not (unlike their North Atlantic counterparts) characterized by the anomalously light values that are indicative of freshwater input during Heinrich Stadials^{12,13} (Fig. 2e). This discrepancy can potentially be attributed to the subsurface habitat of the studied foraminifera¹², highlighting the need for a more suitable proxy for recording surface-water isotope conditions in the North Pacific. Here we provide a roughly 50,000-year-long $\delta^{18}\text{O}$ record from the open northeastern North Pacific reconstructed from diatoms ($\delta^{18}\text{O}_{\text{diat.}}$), which are unicellular algae with siliceous shells that are bound to the surface-water layer. In combination with a water-isotope-enabled, fully coupled climate model⁶, we are able to conduct a direct data–model comparison regarding $\delta^{18}\text{O}$ changes, allowing a fresh perspective on the dynamic link between the climate histories of the North Atlantic and North Pacific during Heinrich Stadials.

Our $\delta^{18}\text{O}_{\text{diat.}}$ record was obtained from kasten core SO202-27-6 (30 cm × 30 cm), recovered in the catchment area of the CIS (54.3° N, 149.6° W; water depth 2,919 m; Fig. 1). The record is characterized by two prominent $\delta^{18}\text{O}_{\text{diat.}}$ minima with transient decreases in $\delta^{18}\text{O}_{\text{diat.}}$ of around 2‰–3‰—one during late Marine Isotope Stage (MIS) 2 (ref. ¹²), and one during MIS3 (Fig. 2b). The two minima are observed with similar magnitude in the reconstructed surface seawater $\delta^{18}\text{O}$ ($\delta^{18}\text{O}_{\text{sw}}$) (Fig. 2c), obtained after correcting the $\delta^{18}\text{O}_{\text{diat.}}$ record for global ice volume and temperature (see Methods). The age of the younger $\delta^{18}\text{O}_{\text{sw}}$ minimum is well constrained, and coincides with Heinrich Stadial 1. The older surface $\delta^{18}\text{O}_{\text{sw}}$ minimum can be aligned to Heinrich Stadial 4 (Methods).

The surface $\delta^{18}\text{O}_{\text{sw}}$ signal is influenced by several factors, including ocean advection and the input of meteoric water. In the glacial North Pacific region, the meteoric input can be attributed mainly to in situ precipitation and CIS meltwater. The latter can be ascribed to the melting of icebergs—that is, when the CIS had a marine grounding line—or to flooding events related to the drainage of (sub)glacial lakes. To simulate climate responses in the North Pacific during Heinrich Stadial 1, we conducted two types of hosing experiments under the conditions of the Last Glacial Maximum (LGM; ref. ¹⁴) by imposing freshwater flux either into the North Atlantic alone (LGM_NA) or additionally into the North Pacific (LGM_NA + NP) (Methods and Extended Data Table 1). To represent $\delta^{18}\text{O}_{\text{sw}}$ changes caused by freshwater input to the North Atlantic and northeastern North Pacific, we used a uniform $\delta^{18}\text{O}$ value of -30‰ for both the LGM_NA and the LGM_NA + NP hosing components, corresponding to the average composition of the LIS (see Methods). This isotopic value enables us to quantify the smallest amount of freshwater needed to account for the reconstructed $\delta^{18}\text{O}$ variations in our record (Methods).

¹Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany. ²MARUM—Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. ³Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Potsdam, Germany. *e-mail: edith.maier@awi.de; xu.zhang@awi.de

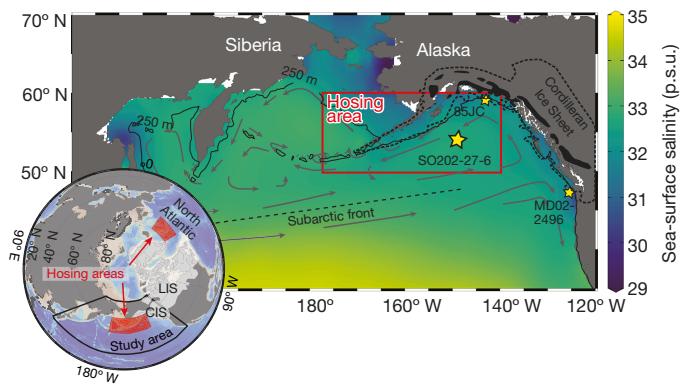


Fig. 1 | Study area. Inset, our study area in the North Pacific, and the areas in the North Atlantic and North Pacific that we chose for freshwater hosing in our palaeoclimate models. The shaded white areas represent the extents of the LIS and CIS during the LGM (see Extended Data Fig. 3). Main image, modern North Pacific sea-surface salinity³⁰ and the northeastern North Pacific hosing area (red rectangle). Arrows represent the modern surface water circulation. Also shown are the locations of cores SO202-27-6 (large yellow star), MD02-2496 and EW0408-85JC (small yellow stars); the modern extent of Cordilleran glaciers (black areas); and the extent of the CIS during the LGM (see Extended Data Fig. 3) (black dashed line). Northern Hemisphere and sea-surface-salinity maps were created using Ocean Data View (see Extended Data Fig. 3).

In LGM_NA, the subarctic northeastern North Pacific is characterized by an increase in surface $\delta^{18}\text{O}_{\text{sw}}$ (Fig. 3d) resulting from enhanced advection of (sub)tropical, $\delta^{18}\text{O}_{\text{sw}}$ -enriched surface waters, which substantially outweighs the counteracting effects of the increased $\delta^{18}\text{O}_{\text{sw}}$ -depleted precipitation (Extended Data Fig. 1a, b). This effect is related to the low glacial sea levels (which were about 120 m lower than at present)¹⁵ and a closed Bering Strait, which prevented the inflow of fresh North Atlantic surface waters to the North Pacific during Heinrich Stadials 1–4 (ref. ¹⁶). Therefore, a substantial freshwater flux (around 0.1 sverdrups, Sv) from the CIS into the North Pacific must be invoked in the model to explain the observed decrease in surface $\delta^{18}\text{O}_{\text{sw}}$ (Figs. 2c, 3i). Indeed, enhanced IRD abundances in glacial deposits until Heinrich Stadial 1 (Fig. 2a)—including pebble-to-cobble-sized dropstones in the intervals assigned to Heinrich Stadials 2 and 4 (Extended Data Fig. 2a)—suggest the existence of a marine-based CIS, that is, a CIS with a grounding line in the North Pacific, during most of the last glacial period. Furthermore, these enhanced IRD abundances indicate that the freshwater flux from melting icebergs reached the open northeastern North Pacific during Heinrich Stadials. Besides iceberg melting, flooding events from (sub)glacial lakes^{17–19} could have provided additional freshwater (Methods).

In contrast to the $\delta^{18}\text{O}_{\text{diat.}}$ record, the $\delta^{18}\text{O}_{\text{pl.foram.}}$ record from the same open ocean core—derived from sinistral *Neogloboquadrina pachyderma* specimens—shows slightly elevated values during Heinrich Stadials 1 and 4 (Fig. 2e), indicating a local cooling and/or enrichment of $\delta^{18}\text{O}_{\text{sw}}$. Sinistral *N. pachyderma* are subsurface dwellers and respond to the depth of the pycnocline, which is at roughly 150 m at the study site (see Methods). In both hosing experiments, simulated subsurface $\delta^{18}\text{O}_{\text{pl.foram.}}$ appears to be in general agreement with observed changes in sinistral *N. pachyderma* $\delta^{18}\text{O}_{\text{pl.foram.}}$ values (Fig. 3e, j). This can probably be attributed to a coherent subsurface cooling in both experiments (Fig. 3b, g), given their contrasting responses in subsurface $\delta^{18}\text{O}_{\text{sw}}$ —namely a depletion in LGM_NA+NP but enrichment in LGM_NA (Extended Data Fig. 1d, h). Like our open-ocean subsurface record, surface $\delta^{18}\text{O}_{\text{sw}}$ records reconstructed from the coastal northeastern North Pacific, based on $\delta^{18}\text{O}_{\text{pl.foram.}}$ data from *Globigerina bulloides*, do not show distinctly depleted $\delta^{18}\text{O}_{\text{sw}}$ values during times of elevated IRD deposition^{13,20,21} (Fig. 4d, e, g). This finding could be related to the proximity of the coastal sites to the CIS margin, which resulted in only a small difference between the local background $\delta^{18}\text{O}_{\text{sw}}$ values and

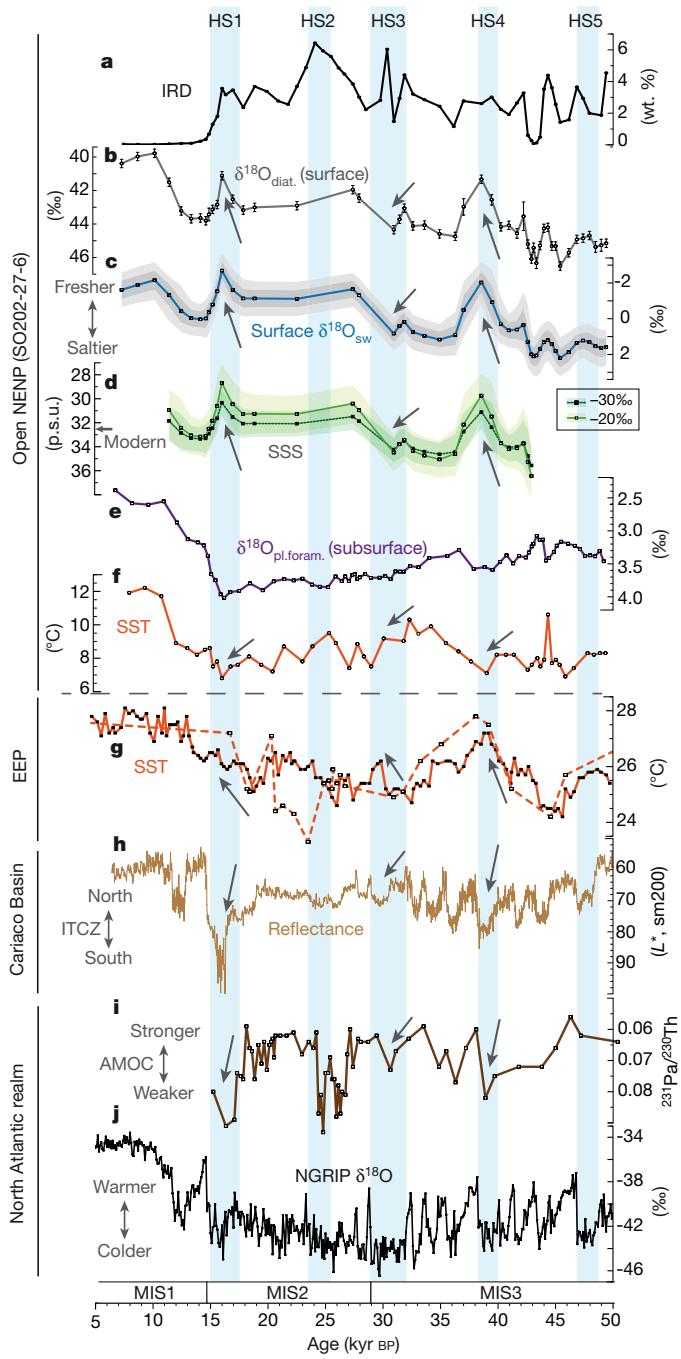


Fig. 2 | Proxy data from the North Pacific and North Atlantic (50 kyr to 5 kyr BP). a–f, Data from northeastern North Pacific core SO202-27-6 (in b, e and f, data for the past 25 kyr BP are from ref. ¹²). a, Ice-raftered debris. b, $\delta^{18}\text{O}_{\text{diat.}}$ data; error bars show the errors of replicate analyses or the long-term reproducibility of standards (1σ). c, Surface $\delta^{18}\text{O}_{\text{sw}}$; dark grey and light grey envelopes show 68% and 95% confidence intervals, respectively. d, Sea-surface salinity calculated from surface $\delta^{18}\text{O}_{\text{sw}}$; green envelopes show 95% confidence intervals, assuming a CIS meltwater $\delta^{18}\text{O}$ of $-20\text{\textperthousand}$ (light green) or $-30\text{\textperthousand}$ (dark green). e, Subsurface $\delta^{18}\text{O}_{\text{pl.foram.}}$ data from sinistral *N. pachyderma*. f, Alkenone-based SSTs. g, Alkenone-based (solid line) and magnesium/calcium-based (dashed line) SSTs (from the eastern Equatorial Pacific, core MD02-2529; ref. ²⁵). h, Sediment total reflectance (from the Cariaco Basin; ref. ²⁴). L^* , lightness; sm200, 200-point running mean. i, $^{231}\text{Pa}/^{230}\text{Th}$ ratio (Ocean Drilling Program (ODP) site 1063; ref. ³). j, NGRIP $\delta^{18}\text{O}$ record⁷. EEP, eastern Equatorial Pacific; HS, Heinrich Stadial; ITCZ, Intertropical Convergence Zone. Arrows indicate the direction of proxy changes during Heinrich Stadials 1, 3 and 4.

the CIS freshwater $\delta^{18}\text{O}$ value (ref. ²⁰). However, the surface $\delta^{18}\text{O}_{\text{sw}}$ at our site (today around $-0.5\text{\textperthousand}$; ref. ²²) differed substantially from the CIS freshwater $\delta^{18}\text{O}$, so our surface-confined $\delta^{18}\text{O}_{\text{diat.}}$ record probably

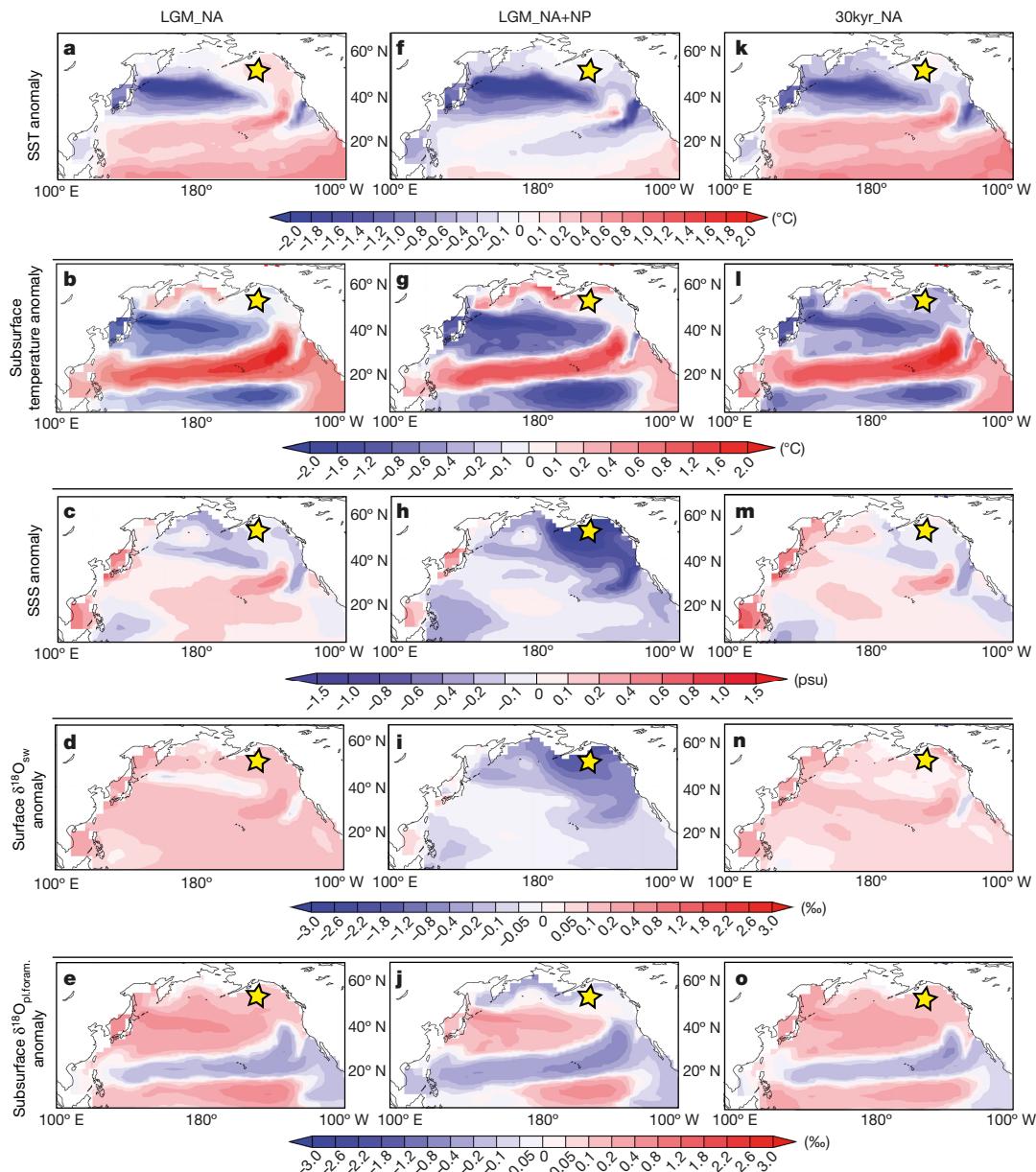


Fig. 3 | Results of freshwater hosing experiments LGM_NA, LGM_NA+NP and 30kyr_NA. Model results are presented as anomalies between the hosing simulations and the LGM state (see Methods). Left, results from LGM_NA. Middle, results from LGM_NA+NP. Right, results

follows the $\delta^{18}\text{O}_{\text{sw}}$ changes associated with freshwater discharges. Given assumed isotopic values of $-20\text{\textperthousand}$ to $-30\text{\textperthousand}$ for glacial Cordilleran ice (Methods), decreases of around 2\textperthousand – 3\textperthousand in surface $\delta^{18}\text{O}_{\text{sw}}$ correspond to decreases in sea-surface salinity (SSS) of around 2–4 practical salinity units (p.s.u.; Fig. 2d and Extended Data Fig. 3). Such decreases are consistent with our LGM_NA+NP results (which show a simulated decrease in SSS of around 2 p.s.u., and in surface $\delta^{18}\text{O}_{\text{sw}}$ of about 2\textperthousand ; Fig. 3h, i). The influence of precipitation changes on SSS is probably minor, given that total precipitation changes by less than 5 mm per month in our LGM_NA+NP experiment, and that the simulated precipitation $\delta^{18}\text{O}$ values decrease only slightly (less than 1\textperthousand ; Extended Data Fig. 1e, f).

The close temporal correlation of CIS freshwater events to Heinrich Stadials indicates a potential dynamic link of meltwater events between the North Atlantic and North Pacific. External forcing of Northern Hemisphere summer insolation is thought to have initially triggered LIS retreat during glacial terminations²³, and might also have driven

from 30kyr_NA. **a, f, k**, SST anomalies. **b, g, l**, Subsurface temperature anomalies (120–180 m depth). **c, h, m**, SSS anomalies. **d, i, n**, Surface $\delta^{18}\text{O}_{\text{sw}}$ anomalies. **e, j, o**, Subsurface $\delta^{18}\text{O}_{\text{plforam}}$ anomalies (150 m). The yellow star marks the location of the studied core SO202-27-6.

CIS freshwater discharge during Heinrich Stadial 1 (ref. ¹²). However, given that such insolation forcing was limited during Heinrich Stadial 4 (Extended Data Fig. 2g), alternative trigger mechanisms should be considered. We propose that the observed recurring CIS meltwater events can be attributed to a weakened AMOC, which leads to positive feedbacks in the northeastern North Pacific in the presence of a marine-based CIS, through interactions between low and high latitudes and between the ocean and the atmosphere.

In our LGM_NA experiment, the weakened AMOC reduces meridional heat transport to the northern high latitudes, resulting in a southward shift of the Intertropical Convergence Zone²⁴ (Fig. 2h, i) and increased (sub)surface temperatures in the eastern Equatorial Pacific^{25,26} (Figs. 2g and 3a, b). As a consequence, rainfall in the western Equatorial Pacific decreased, ultimately strengthening the Aleutian Low pressure system²⁷ (see Methods and Extended Data Fig. 4a). This led to increased poleward transport of (sub)tropical waters under the southeastern flank of the enhanced Aleutian

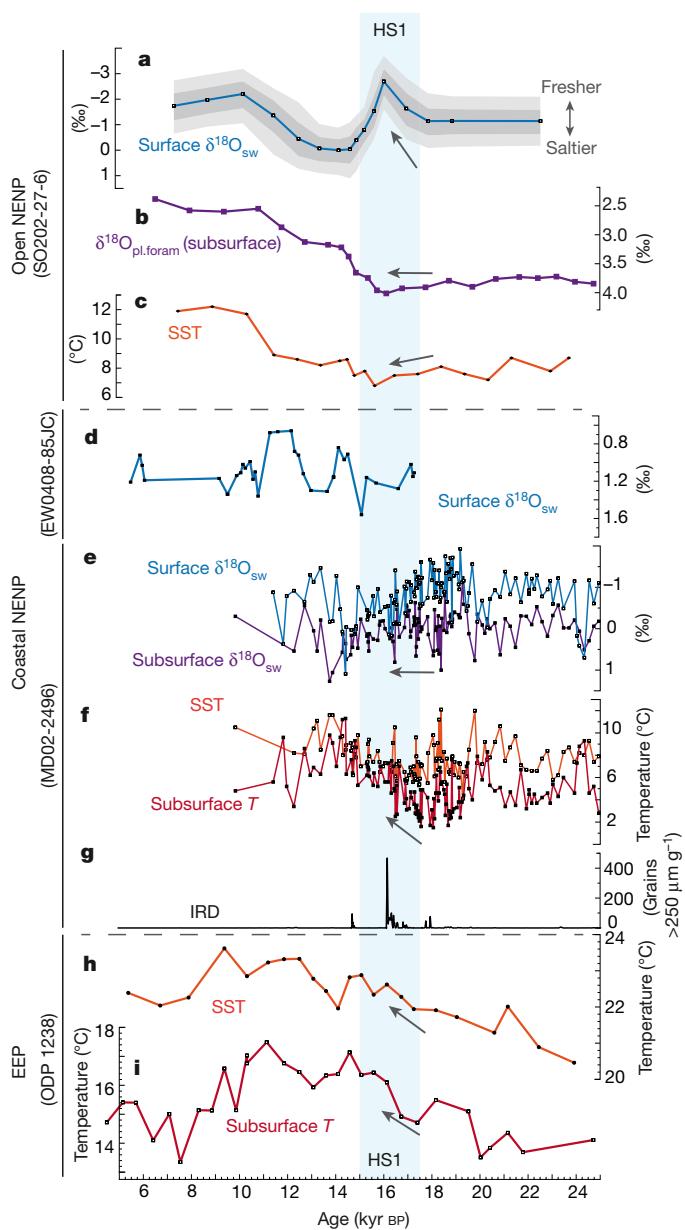


Fig. 4 | Proxy data from the eastern Equatorial Pacific (EEP) and northeastern North Pacific (NENP). **a–c**, Open-ocean NENP (core SO202-27-6). **a**, Surface $\delta^{18}\text{O}_{\text{sw}}$ with dark grey and light grey envelopes indicating the 68% and 95% confidence intervals, respectively, including age and analytical uncertainties. **b**, Subsurface sinistral *N. pachyderma* $\delta^{18}\text{O}_{\text{pl. foram}}$ (ref. ¹²). **c**, Alkenone-based SSTs¹². **d–f**, Coastal NENP. **d**, Surface $\delta^{18}\text{O}_{\text{sw}}$ (from core EW0408-85JC; ref. ²¹). **e**, (Sub)surface $\delta^{18}\text{O}_{\text{sw}}$ (ref. ²⁰). **f**, (Sub)surface temperature (*T*; ref. ²⁰) from core MD02-2496. **g**, IRD¹¹ from core MD02-2496. **h, i**, EEP. **h**, SSTs (magnesium/calcium-based, from *Globigerinoides sacculifer*) from ODP site 1238 (ref. ²⁵). **i**, Subsurface temperature (magnesium/calcium-based, from *Neogloboquadrina dutertrei*) from ODP site 1238 (ref. ²⁵). Arrows indicate the direction of proxy changes during Heinrich Stadials 1, 3 and 4.

Low, causing subsurface warming along the coastal northeastern North Pacific (Figs. 3b and 4f). The subsurface warming resulted in increased basal melting/calving of the marine-based CIS, leading to freshwater input to the northeastern North Pacific—a similar feedback to that proposed for glacial discharge events from the LIS²⁸ and the Antarctic ice sheets²⁹. Our LGM_NA+NP results show weakened vertical mixing in response to surface water freshening, causing surface cooling and subsurface warming (Figs. 2d, f and 3f–h), which act as a local positive feedback mechanism to further accelerate the release of CIS meltwater.

This proposed mechanism links climate fluctuations observed in the North Atlantic, eastern Equatorial Pacific and northeastern North Pacific during Heinrich Stadials 1 and 4. During Heinrich Stadial 3, however, the increases in SSS and surface $\delta^{18}\text{O}_{\text{sw}}$ at our site do not indicate increased CIS meltwater discharge (Fig. 2c, d). A robust data evaluation for Heinrich Stadial 3 is precluded by the absence of $\delta^{18}\text{O}_{\text{diat}}$ data for the time of the maximum Heinrich Stadial 3 IRD abundance, as a consequence of the low biogenic opal content of less than 5% in the sediments (as for the sediments corresponding to Heinrich Stadial 2). Nevertheless, to test the dynamic link between the North Atlantic and North Pacific during this stadial, we performed an additional North Atlantic hosing experiment under conditions of 30,000 years ago (30kyr_NA; Extended Data Table 1). As for LGM_NA, 30kyr_NA shows warming in the eastern Equatorial Pacific and an increased Aleutian Low (Fig. 3k, l and Extended Data Fig. 4b), indicating that the first part of the dynamic link—that is, the teleconnection between the North Atlantic and eastern Equatorial Pacific—also works during Heinrich Stadial 3. However, the simulated (sub)surface cooling, surface $\delta^{18}\text{O}_{\text{sw}}$ enrichment and SSS increase at our site (Fig. 3k–n), which match our proxy data (Fig. 2c, d, f), indicate that the warm and salty (sub)tropical water masses cooled down before reaching the coastal northeastern North Pacific. Given that Heinrich Stadial 3 was the coldest period of the past 50,000 years in the northern high latitudes⁷, it seems that the enhanced Northern Hemisphere cooling under 30-kyr orbital forcing supersedes the warming effect from the subtropics, preventing massive CIS meltwater events during Heinrich Stadial 3 (Methods). Therefore, background cooling in the northern high latitudes acts as a critical negative feedback on the collapse of marine-based CIS ice, modulating the dynamic link between the North Atlantic and North Pacific during Heinrich Stadials.

The results of our data–model comparison provide compelling evidence that, during North Atlantic cold stadials characterizing the past 50,000 years, perturbations to the AMOC could have been teleconnected to the northeastern North Pacific region, triggering freshwater discharge events via interactions between low and high latitudes and between oceans and the atmosphere. Until now, such North Pacific freshwater input events have not been considered as standard forcing components in glacial climate simulations; the incorporation of this freshwater forcing scenario provides a new basis for research that could reconcile the discrepancies within proxy data regarding the responses of North Pacific ocean circulation to AMOC changes. For example, because of the limitations of age-model constraints in the North Pacific (related to poor knowledge of palaeoreservoir ages), it is difficult to assess the lead–lag relationship of North Pacific meltwater events with changes in the AMOC using proxy data. However, on the basis of a North Pacific hosing experiment with LGM boundary conditions (LGM_NP; Extended Data Table 1), it seems that North Pacific meltwater discharge alone leads to subsurface cooling in the North Atlantic (Extended Data Fig. 5c), acting as an unlikely trigger of ice-surging events in the North Atlantic²⁸ and related AMOC changes (see Methods). Given that glacial meltwater events are closely associated with ice-sheet dynamics, climate models that incorporate interactive ice-sheet dynamics together with high-resolution proxy records from the open northeastern North Pacific are highly desirable to further assess the proposed dynamic linkages between the North Atlantic and North Pacific, as well as local feedbacks within the North Pacific.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0276-y>.

Received: 26 July 2017; Accepted: 14 May 2018;
Published online 11 July 2018.

1. Maslin, M. A., Shackleton, N. J. & Pflaumann, U. Surface water temperature, salinity, and density changes in the northeast Atlantic during the last 45,000 years: Heinrich events, deep water formation, and climatic rebounds. *Paleoceanography* **10**, 527–544 (1995).

2. Kageyama, M. et al. Climatic impacts of fresh water hosing under Last Glacial Maximum conditions: a multi-model study. *Clim. Past* **9**, 935–953 (2013).
3. Böhm, E. et al. Strong and deep Atlantic meridional overturning circulation during the last glacial cycle. *Nature* **517**, 73–76 (2015).
4. Praetorius, S. & Mix, A. Synchronisation of North Pacific and Greenland climates preceded abrupt deglacial warming. *Science* **345**, 444–448 (2014).
5. Menviel, L., England, M. H., Meissner, K. J., Mouchet, A. & Yu, J. Atlantic-Pacific seesaw and its role in outgassing CO₂ during Heinrich events. *Paleoceanography* **29**, 58–70 (2014).
6. Werner, M. et al. Glacial-interglacial changes in H₂¹⁸O, HDO and deuterium excess—results from the fully coupled ECHAM5/MPI-OM Earth system model. *Geosci. Model Dev.* **9**, 647–670 (2016).
7. North Greenland Ice Core Project members. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* **431**, 147–151 (2004).
8. Henry, L. G. et al. North Atlantic ocean circulation and abrupt climate changes during the last glaciation. *Science* **353**, 470–474 (2016).
9. Chikamoto, M. O. et al. Variability in North Pacific intermediate and deep water ventilation during Heinrich events in two coupled climate models. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **61–64**, 114–126 (2012).
10. Seguinot, J., Rogozhina, I., Stroeven, A. P., Margold, M. & Kleman, J. Numerical simulations of the Cordilleran ice sheet through the last glacial cycle. *Cryosphere Discuss.* **9**, 4147–4203 (2015).
11. Hendy, I. L. & Cosma, T. Vulnerability of the Cordilleran Ice Sheet to iceberg calving during late Quaternary rapid climate change events. *Paleoceanography* **23**, PA2101 (2008).
12. Maier, E. et al. Deglacial subarctic Pacific surface water hydrography and nutrient dynamics and links to North Atlantic climate variability and atmospheric CO₂. *Paleoceanography* **30**, 949–968 (2015).
13. Gebhardt, H. et al. Paleonutrient and productivity records from the subarctic North Pacific for Pleistocene glacial terminations I to V. *Paleoceanography* **23**, PA4212 (2008).
14. Zhang, X., Lohmann, G., Knorr, G. & Xu, X. Different ocean states and transient characteristics in Last Glacial Maximum simulations and implications for deglaciation. *Clim. Past* **9**, 2319–2333 (2013).
15. Siddall, M. et al. Sea-level fluctuations during the last glacial cycle. *Nature* **423**, 853–858 (2003).
16. Hu, A. et al. Influence of Bering Strait flow and North Atlantic circulation on glacial sea-level changes. *Nat. Geosci.* **3**, 118–121 (2010).
17. Wiedmer, M., Montgomery, D. R., Gillespie, A. R. & Greenberg, H. Late Quaternary megafloods from Glacial Lake Atna, Southcentral Alaska, U.S.A. *Quat. Res.* **73**, 413–424 (2010).
18. Benito, G. & O'Connor, J. E. Number and size of last-glacial Missoula floods in the Columbia River valley between the Pasco Basin, Washington, and Portland, Oregon. *Geol. Soc. Am. Bull.* **115**, 624–638 (2003).
19. Livingstone, S. J., Clark, C. D. & Tarasov, L. Modelling North American palaeo-subglacial lakes and their meltwater drainage pathways. *Earth Planet. Sci. Lett.* **375**, 13–33 (2013).
20. Taylor, M. A., Hendy, I. L. & Pak, D. K. Deglacial ocean warming and marine margin retreat of the Cordilleran Ice Sheet. *Earth Planet. Sci. Lett.* **403**, 89–98 (2014).
21. Praetorius, S. et al. North Pacific deglacial hypoxic events linked to abrupt climate warming. *Nature* **527**, 362–366 (2015).
22. Kipphut, G. W. Glacial meltwater input to the Alaska Coastal Current: evidence from oxygen isotope measurements. *J. Geophys. Res.* **95**, 5177–5181 (1990).
23. Cheng, H. et al. Ice age terminations. *Science* **326**, 248–252 (2009).
24. Deplazes, G. et al. Links between tropical rainfall and North Atlantic climate during the last glacial period. *Nat. Geosci.* **6**, 213–217 (2013).
25. Martínez-Botí, M. A. et al. Boron isotope evidence for oceanic carbon dioxide leakage during the last deglaciation. *Nature* **518**, 219–222 (2015).
26. Leduc, G. et al. Moisture transport across Central America as a positive feedback on abrupt climatic changes. *Nature* **445**, 908–911 (2007).
27. Okumura, Y. M., Deser, C., Hu, A., Timmermann, A. & Xie, S.-P. North Pacific climate response to freshwater forcing in the subarctic North Atlantic: oceanic and atmospheric pathways. *J. Clim.* **22**, 1424–1445 (2009).
28. Marcott, S. A. et al. Ice-shelf collapse from subsurface warming as a trigger for Heinrich Events. *Proc. Natl. Acad. Sci. USA* **108**, 13415–13419 (2011).
29. Weber, M. E. et al. Millennial-scale variability in Antarctic ice-sheet discharge during the last deglaciation. *Nature* **510**, 134–138 (2014).
30. Antonov, J. I. et al. *World Ocean Atlas 2009*, Vol. 2: Salinity. (NOAA Atlas NESDIS 69, US Gov. Print. Off., Washington DC, 2010).

Acknowledgements This work was largely part of the Innovative NOrth Pacific Experiment (INOPEX), funded by the Bundesministerium für Bildung und Forschung. We also acknowledge funding by the Helmholtz Postdoc program (PD-301; to X.Z.), as well as Helmholtz funding through the Polar Regions and Coasts in the Changing Earth System (PACES) program of the Alfred Wegener Institute. Funding from the Qingdao National Laboratory for Marine Science and Technology (QNL201703) is also acknowledged. We thank U. Böttjer, B. Glückselig and R. Cordelair for the thorough purification of diatom samples for isotope analyses; M. Warnkross for picking planktic foraminifera for stable-isotope analysis and radiocarbon dating; S. Steph and A. Mackensen for performing the foraminiferal oxygen-isotope analysis; and G. Knorr for helpful discussions.

Reviewer information *Nature* thanks S. Dee, A. Hu, K. Thirumalai and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions E.M. and X.Z. designed the study and wrote the manuscript with contributions from A.A., R.G. and G.L. E.M. performed the diatom isotope measurements with support from B.C. and H.M. X.Z. designed the model experiments and performed simulations with support from M.W. and G.L. E.M. constructed the age model and S.M. carried out the proxy uncertainty modelling. E.M. performed the contamination analysis of diatom samples. M.M. and R.S. contributed alkenone-based sea-surface temperatures (SSTs), and J.R. the diatom composition of the isotope samples. All authors contributed to the final version of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0276-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to E.M. or X.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Chronology. The chronology of core SO202-27-6 is based on 12 planktic ^{14}C ages, in combination with 15 additional age control points. The planktic foraminifera for planktic ^{14}C ages were picked from the 125–250- μm fraction and dated by accelerator mass spectrometry (AMS) at the National Ocean Science AMS facility (NOSAMS) at Woods Hole Oceanographic Institution (Extended Data Table 2). The 15 additional age control points, in between the planktic ^{14}C ages and below 243.5 cm, were obtained through two different approaches. First, in the upper-core section (0–91 cm), the five additional age control points (LS-1 to LS-5) correspond to the calibrated ages determined in ref. ¹², which are ^{14}C plateau boundaries determined via proxy correlation to the high-resolution ^{14}C record of core MD02-2489 (ref. ¹²). Second, the ten age control points in the lower part of the core (91–289 cm; NG-1 to NG-10) were determined by visual correlation between iron intensity and the NGRIP dust record³¹ (Extended Data Table 2; Extended Data Fig. 2), assuming in-phase behaviour between both parameters. This assumption is reasonable because of the in-phase relationship between NGRIP dust and subarctic Pacific eolian dust³², and between eolian dust and the iron intensity of northwestern North Pacific core SO202-07-6, over the last glacial–interglacial transition (Extended Data Fig. 6). The correlation between iron intensity and NGRIP dust record results in overall in-phase behaviour between increased calcium concentrations in the subarctic northeastern North Pacific and warm periods in the North Atlantic region, consistent with previous MIS3 age models from the northeastern North Pacific^{33,34}.

We used the R script BACON (ref. ³⁵) version 2.2 and the Marine13 calibration curve³⁶ to model down-core calendar age uncertainty. For the upper section of core SO202-27-6 (0–91 cm), we applied additional local reservoir ages (ΔR) of 40 ± 173 years (1σ) and 150 ± 185 years (1σ), consistent with the reservoir ages determined for nearby core MD02-2489 (ref. ³⁷), which was correlated to the upper part of core SO202-27-6 (0–91 cm) via proxy data¹². For the lower-core section (91–289 cm) we applied a local reservoir age (ΔR) of 710 ± 202 years (1σ), as determined for ^{14}C -plateau IV for core MD02-2489 (ref. ³⁷), because no information is available on local reservoir ages and reservoir age changes during MIS3. Down-core calendar age distributions were modelled using BACON's default settings and a Student's t distribution (shape parameter $t.a. = 20$, scale parameter $t.b. = 21$) (Extended Data Fig. 7). From 10,000 age–depth realizations generated with BACON, we calculated the median age and the 95% confidence intervals at 1-cm resolution.

X-ray fluorescence measurements. The relative elemental compositions (in counts per second; c.p.s.) of core SO202-27-6 and core SO202-07-6 were measured at 1-cm resolution using an AvaaTech X-ray fluorescence (XRF) core scanner located at the Alfred Wegener Institute, Bremerhaven, Germany. The elements iron and calcium used here for core chronology were obtained from scans performed at 1 mA, with a tube voltage of 10 kV and a counting time of 30 seconds.

Diatom samples for oxygen-isotope analysis. Diatom valves for oxygen-isotope analysis (91–289 cm) were extracted from bulk sediment samples obtained every 5 cm (2-cm-thick slices; 225 cm^3) from kasten core SO202-27-6. Diatom samples (100–125- μm fraction) were purified as in refs ^{12,38}. Briefly, bulk samples were liberated from carbonates and non-diatom silicates using a combination of physical and chemical treatments, including sonication and heavy liquid separation. Diatom assemblages were determined from microscopic slides, prepared before the sonication procedure, following the diatom taxonomy of refs ^{39,40}. Purified diatom samples were dominated by *Coscinodiscus* species ($99.7 \pm 1.2\%$), with a large contribution of *C. marginatus* ($92.7 \pm 9.6\%$) and a minor contribution of *C. oculus-iridis* ($7.0 \pm 9.5\%$) (Extended Data Fig. 8a). We checked the purified diatom samples for non-biogenic silicate contamination using energy-dispersive X-ray spectrometry (EDS) on subsamples of all purified diatom samples, and checked the EDS results by additional measurements using inductively coupled plasma optical emission spectrometry on six diatom samples (see also refs ^{12,38}). Aluminium oxide was used as a tracer for non-biogenic silicates, and mass balance corrections were applied⁴¹, using two different $\delta^{18}\text{O}$ values ($+2\%$ and $+30.00\%$) for the contamination, corresponding to the isotopic range of non-biogenic silicates^{42,43}. Contamination of all diatom samples from the last glacial period is less than 4%, except in the case of three samples from early MIS2 (Extended Data Fig. 8b), indicating a high purity of diatom samples. Mass-balance-corrected isotopic curves still show pronounced $\delta^{18}\text{O}_{\text{diat.}}$ minima during Heinrich Stadials 1 and 4 (Extended Data Fig. 8c), showing that the $\delta^{18}\text{O}_{\text{diat.}}$ minima are not the result of silicate contamination.

Diatom oxygen-isotope measurements. We used about 1.5–2.0 mg of purified material from SO202-27-6 diatom samples (91–289 cm) in order to measure the $\delta^{18}\text{O}_{\text{diat.}}$ composition using laser fluorination and a PDZ Europa 2020 mass spectrometer⁴⁴. Values are reported in the common δ notation versus Vienna Standard Mean Ocean Water (V-SMOW), using the laboratory diatom standards PS1772-8_{bsis} (marine) and BFC (lacustrine) calibrated against the International Atomic Energy Agency (IAEA) reference quartz standard National Bureau of Standards-28 (NBS-28). Analytical precision—determined by repeated analyses of PS1772-8_{bsis} (two batches) and BFC over the periods when the samples were measured—was

better than 0.25% (1σ), in line with published long-term reproducibility from this instrumentation⁴⁵. (PS1772-8_{bsis} (batch used for the interlaboratory comparison⁴⁵): precision $43.49\% \pm 0.16\%$, $n = 40$; PS1772-8_{bsis} (subsequent batch): precision $44.15\% \pm 0.19\%$, $n = 22$; BFC: precision $28.81\% \pm 0.24\%$, $n = 10$.) Diatom samples were measured at least twice when enough purified material was available. Our record extends the $\delta^{18}\text{O}_{\text{diat.}}$ record (0–91 cm) published in ref. ¹².

Foraminifera oxygen-isotope measurements. $\delta^{18}\text{O}$ measurements on sinistral *N. pachyderma* from core SO202-27-6 (91–289 cm) were carried out using a MAT 251 mass spectrometer directly coupled to an automated carbonate preparation device (Kiel I), and calibrated via the National Institute of Standards and Technology-19 (NIST-19) international standard to the Pee Dee Belemnite (PDB) scale. Sinistral *N. pachyderma* were picked from the 125–250 μm and the 315–400 μm fractions. All isotope values are given in δ notation versus Vienna-PDB (V-PDB). The precision of the measurements at 1σ , determined by repeated measurements of the internal Solnhofen limestone over a one-year period, was better than 0.08% . This record extends the sinistral *N. pachyderma* $\delta^{18}\text{O}$ record (0–91 cm) of ref. ¹².

IRD calculation. As an indicator of the abundance of IRD, we calculated the weight percentage of lithic and mineral grains of the >250- μm -to-2-mm fraction (medium-to-coarse sand), in accordance with previous IRD studies in the Gulf of Alaska^{46,47}. We separated the lithic/mineral grains from the biogenic silicates by performing a heavy liquid separation (density = 2.2–2.3 g cm^{-3}) after organic and carbonate removal using hydrogen peroxide and hydrochloric acid. We regard the removal of carbonates as reasonable given that previous studies of sediments from the Gulf of Alaska showed that IRD in the open Gulf of Alaska generally consists of silicate minerals as well as siliciclastic, volcanic and metamorphic rock fragments⁴⁶. We then sieved the heavy fraction at 250 μm and 2 mm (the light fraction was further cleaned for diatom isotope analysis) and normalized the weight of the >250- μm -to-2-mm fraction to the dry bulk sample weight. Dropstones (IRD larger than 2 mm) are not included in the calculation.

Alkenone-based SSTs. We determined alkenone-based SSTs from samples of core SO202-27-6 (91–289 cm) through gas chromatography (GC) and GC/mass spectrometry⁴⁸. We determined the SSTs using the calibration of ref. ⁴⁹, providing reasonable summer SST estimates for the (sub)arctic Pacific⁴⁸. The standard error of the calibration is 1.5°C . The total analytical error calculated from replicate analyses of an external alkenone standard (extracted from *Emiliana huxleyi* (EHUX) cultures with known growth temperatures) is less than 0.4°C . Our SST record extends the record (0–91 cm) of ref. ¹².

Calculation and error analysis of $\delta^{18}\text{O}_{\text{sw}}$ and SSS. To calculate local surface $\delta^{18}\text{O}_{\text{sw}}$ we generated 10,000 noisy Monte Carlo proxy realizations for the alkenone SSTs and $\delta^{18}\text{O}_{\text{diat.}}$ within the analytical uncertainty. We combined both proxy ensembles with BACON's age ensemble of similar size to obtain 10,000 possible time series for each proxy. All time series were interpolated to the median age obtained for the depth in which $\delta^{18}\text{O}_{\text{diat.}}$ was measured. The resulting SST and $\delta^{18}\text{O}_{\text{diat.}}$ time-series ensembles were used to calculate an ensemble of local surface $\delta^{18}\text{O}_{\text{sw}}$ with the following equation (ref. ⁵⁰):

$$\text{Local surface } \delta^{18}\text{O}_{\text{sw}} = \delta^{18}\text{O}_{\text{diat.}} - 34 - \sqrt{122 - 5\text{SST}} - \text{mean}(\delta^{18}\text{O}_{\text{sw}})$$

with $\delta^{18}\text{O}_{\text{diat.}}$ being the measured diatom $\delta^{18}\text{O}$, SST being the temperature calculated from the alkenone-based SST record from the same core, and $\text{mean}(\delta^{18}\text{O}_{\text{sw}})$ being the mean seawater $\delta^{18}\text{O}$. To account for global ice-volume-related changes in seawater $\delta^{18}\text{O}$, we corrected the resulting ensemble of local surface $\delta^{18}\text{O}_{\text{sw}}$ with a noisy ensemble of mean seawater $\delta^{18}\text{O}$ (from ref. ⁵¹) before calculating the error envelopes. We consider the use of alkenone-based SSTs for the calculation of $\delta^{18}\text{O}_{\text{sw}}$ to be reasonable given that both coccolithophorids (which comprise the alkenone compounds) and diatoms of the genus *Coscinodiscus* (which make the $\delta^{18}\text{O}_{\text{diat.}}$ signal) (Extended Data Fig. 8a) contribute mainly to late-summer/autumn algal blooms in the subarctic Pacific region^{48,52}.

From the ensemble of ice-effect-corrected local surface $\delta^{18}\text{O}_{\text{sw}}$ records, we then calculated three different SSS records, assuming a linear regression between SSS and $\delta^{18}\text{O}_{\text{sw}}$ (Extended Data Fig. 3). We used a high-salinity endmember corresponding to simulated glacial subsurface waters at the study site (LGM control experiment at 150 m) ($\delta^{18}\text{O} = 0.71\%$; salinity = 34.29 p.s.u.), which was introduced to the euphotic zone during autumn/winter mixing. For the low-salinity endmember (salinity = 0), we applied three freshwater sources with different $\delta^{18}\text{O}$ values, which correspond to: (1) the roughly average composition of modern CIS glaciers^{22,53} (-20%); (2) the roughly average composition of the LIS (-30%)⁵⁴; and (3) the average composition of modern precipitation at sea level in the Gulf of Alaska region (-8.6%)⁵⁵. We performed SSS estimations for only the time interval 43–11 kyr ago, which corresponds to the time period when the Bering strait was closed^{16,56}, as in our LGM control experiment¹⁴.

Model description. We used a comprehensive, fully coupled atmosphere–ocean general circulation model (AO-GCM), namely COSMOS (ECHAM5-JSBACH-MPI-OM), for this study. The atmospheric component of this model

(ECHAM5)⁵⁷—complemented by a land-surface component, JSBACH⁵⁸—is used at ‘T31’ resolution (roughly 3.75°), with 19 vertical layers. The ocean model MPI-OM⁵⁹, which includes sea-ice dynamics that is formulated using viscous-plastic rheology⁶⁰, has a resolution of ‘GR30’ (roughly 3°) in the horizontal, with 40 uneven vertical layers. It has been used for a range of transient simulations, including a North Atlantic freshwater hosing simulation under preindustrial and LGM background states¹⁴, and glacial millennial-scale climate variability^{61,62}. To provide a direct comparison of water oxygen isotopes between proxy records and model outputs, we used the water-isotope-enabled version of COSMOS that has been used to simulate preindustrial and LGM distributions of $\delta^{18}\text{O}$, which are broadly consistent with observations⁶.

Design of LGM_NA and LGM_NA+NP hosing experiments. To mimic the freshwater event from the CIS and explore its dynamic link to changes in the AMOC, we conducted two types of hosing experiment under LGM boundary conditions¹⁴. One was the typical North Atlantic hosing experiment, in which a constant freshwater flux of 0.15 Sv was imposed to the North Atlantic (40–55° N, 20–45° W) to mimic the freshwater discharge during Heinrich Stadial 1 (LGM_NA; Extended Data Table 1). The other hosing experiment also included a freshwater flux of 0.1 Sv to the northeastern North Pacific (50–60° N, 143–172° W) to represent rapid CIS retreat during Heinrich Stadial 1 (LGM_NA+NP). The isotopic values of the imposed freshwater in both cases were $-30\text{\textperthousand}$. This value is consistent with the average composition of the LIS⁵⁴, but is also specific to represent the assumed lower limit of isotopic value in the meltwater from the CIS. This set-up helped us to quantify the minimum amount of freshwater that is needed to reproduce the recorded magnitude of $\delta^{18}\text{O}_{\text{sw}}$ changes in the North Pacific, given that on the one hand meltwater from glacial lakes could be characterized by enriched $\delta^{18}\text{O}$ due to evaporative enrichment, and on the other hand such meltwater could be partly diluted before reaching our open-ocean study site. Both simulations were integrated for 800 model years, and the average of the last 100 years was used to represent the corresponding climatology. The reference climate state is the last 100-year average of the LGM simulation⁶. We note that the simulated $\delta^{18}\text{O}_{\text{sw}}$ at our core site does not differ substantially from that in the coastal regions, since our climate model (like many other climate models) is not able to explicitly resolve coastal hydrology.

Subsurface temperature and subsurface $\delta^{18}\text{O}_{\text{sw}}$ anomalies were taken for a water depth of 120–180 m. This corresponds to the water depth of the shelf area during the last glacial, considering the depth of the shelf area along the Pacific northwest coast today (250–300 m) and a maximum sea level lowstand of about -120 m during the last glacial¹⁵. The $\delta^{18}\text{O}_{\text{pl,foram}}$ anomalies were taken from a water depth of 150 m, which corresponds to the depth of the pycnocline at the study site⁶³ and thus to the assumed habitat depth of subsurface sinistral *N. pachyderma* (ref. ⁶⁴). To calculate $\delta^{18}\text{O}_{\text{pl,foram}}$ from simulated subsurface $\delta^{18}\text{O}_{\text{sw}}$ and simulated subsurface temperature, we used the equation for *N. pachyderma* from ref. ⁶⁵.

Evaluation of freshwater sources. The main sources of freshwater to the northeastern North Pacific are freshwater from the CIS and precipitation. Besides iceberg melting, CIS meltwater related to flooding events from glacial lakes—located, for example, in Alaska¹⁷ and at the southern lobe of the CIS¹⁸—and from subglacial lakes beneath the CIS¹⁹ could have influenced our study site. Our SSS reconstruction (see above) shows that, when using the precipitation low-salinity endmember, SSS changes by about 10 p.s.u. during times of most depleted surface $\delta^{18}\text{O}_{\text{sw}}$, which would require a massive increase in precipitation and/or a massive decrease in precipitation $\delta^{18}\text{O}$. However, our LGM_NA results suggest only a small increase in precipitation of about 1–5 mm per day, with a small decrease in precipitation $\delta^{18}\text{O}$ to at most around $0.4\text{\textperthousand}$, in the northeastern North Pacific during Heinrich Stadial 1 (Extended Data Fig. 1a, b). We therefore reject the idea that an increase in precipitation alone could be responsible for the surface $\delta^{18}\text{O}_{\text{sw}}$ minima. Moreover, precipitation contains a less negative $\delta^{18}\text{O}$ value compared with CIS freshwater, and therefore has a much lower effect on surface $\delta^{18}\text{O}_{\text{sw}}$. By contrast, using the ice-sheet endmembers with $\delta^{18}\text{O}$ values of $-20\text{\textperthousand}$ and $-30\text{\textperthousand}$ requires SSS changes of around 2–4 p.s.u. during Heinrich Stadials 1 and 4. Such SSS changes, and the observed magnitude of surface $\delta^{18}\text{O}_{\text{sw}}$ depletion of 2\textperthousand – 3\textperthousand , are supported by our LGM_NA+NP experiment (Fig. 3h, i).

SSTs in the Equatorial Pacific and the Aleutian Low. Tropical SST responses are generally used to explain the enhanced Aleutian Low during North Atlantic cold events²⁷. As a consequence of warming in the eastern Equatorial Pacific, rainfall in the western Pacific decreased by weakening the Walker Circulation and strengthening the Aleutian Low through triggering the Pacific–North American pattern²⁷. To substantiate this dynamic link in our study, and to evaluate the different regional contributions of climatological SST changes to climate responses over the North Pacific between the LGM and Heinrich Stadial 1, we conducted five sensitivity experiments in ECHAM5 (L19/T31), the atmospheric component of our GCM (AGCM) (Extended Data Table 1). In these AGCM experiments, we used LGM boundary conditions (that is, orbital parameters, topography land–sea mask, ice sheet and greenhouse-gas concentrations). The ‘atmospheric LGM’ (A_LGM)

control run in the AGCM was forced by climatology monthly mean SSTs and sea-ice cover from the LGM control experiment of the fully coupled GCM COSMOS; the ‘atmospheric Heinrich Stadial 1’ (A_HS1) control run in the AGCM was forced by SSTs and sea-ice cover from the hosing experiment LGM_NA.

To investigate the individual contributions of SST changes over different basins to Aleutian Low development over the North Pacific during Heinrich Stadial 1, we conducted three sensitivity experiments in which regional SST fields from the experiment LGM_NA were imposed upon the LGM control SST background, such as the Atlantic basin (30° S to 80° N) (A_HS1_Atl), the eastern Equatorial Pacific (180° E to around 70° W, 25° S to 25° N) (A_HS1_EEP), and a combination of the Atlantic and eastern Equatorial Pacific (A_HS1_EEPAtl), similar to ref. ⁶⁶. The atmosphere model was integrated for 50 years for each model experiment, and the last 30 years were taken to calculate climatological fields. Through these AGCM runs, we quantified the contributions of SST changes in the Atlantic and/or Equatorial Pacific to the strength of the Aleutian Low (Extended Data Fig. 4c–f). It is evident that warming in the eastern Equatorial Pacific is crucial for strengthening of the Aleutian Low as the AMOC slows down. In addition, the simulated eastern Equatorial Pacific warming in our fully coupled AO-GCM COSMOS is broadly consistent with glacial North Atlantic hosing experiments of other Palaeoclimate Modelling Intercomparison Project 3 (PMIP3) models². We therefore suggest that warming of the Equatorial Pacific is a key component that bridges North Atlantic cooling to northeastern North Pacific subsurface warming by modulating the strength of the Aleutian Low.

Background cooling and North Pacific subsurface temperature. Subsurface temperature in the northeastern North Pacific is subject to the combined effects of cold-water masses from the northwestern North Pacific and warm-water masses from the (sub)tropical North Pacific. Therefore, it is plausible that enhanced cooling in the northern high latitudes can weaken and even reverse the subsurface warming in the northern North Pacific that is associated with northward advection of (sub)tropical warm-water masses. This will eventually stabilize the marine-based CIS, reducing meltwater input from the CIS. The boundary conditions during Heinrich Stadial 3—resulting from the lower obliquity compared with Heinrich Stadials 1 and 4—appear to favour the enhanced cooling in the northern high latitudes, ameliorating the collapse of marine-based ice along the North Pacific coastlines. To test this hypothesis, we conducted a North Atlantic hosing experiment (0.15 Sv) under the boundary conditions of 30 kyr ago (30kyr_NA) to mimic the freshwater flux to the North Atlantic during Heinrich Stadial 3 (Extended Data Table 1). Given the uncertainties of sea-level reconstructions^{67,68} and the similar greenhouse gases at 21 kyr and 30 kyr before present (ref. ⁶⁹), we specify the 30-kyr-ago orbital parameters⁷⁰ to our LGM experiment to represent the climate of 30 kyr ago. This also helps us to quantify the additional contribution of the low obliquity to the high-latitude cooling under LGM conditions. As shown in Fig. 3, the 30kyr_NA results substantiate our hypothesis that additional high-latitude cooling associated with low obliquity during Heinrich Stadial 3 causes the subsurface cooling in the northeastern North Pacific, reducing the retreat of the marine-based CIS.

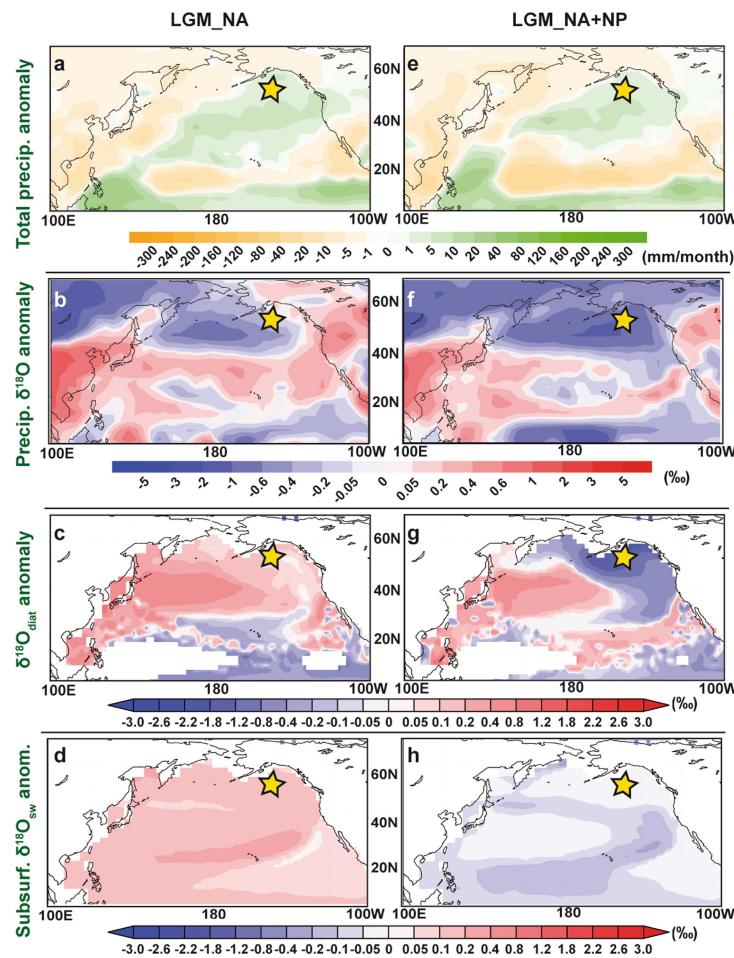
It is worth noting that other factors (the strength of the AMOC itself, ice-sheet configurations, greenhouse gases, and so on) can also determine the background climate. In the above hosing experiments, the ice-sheet configuration and greenhouse gases are identical to their LGM levels, which already lead to a cold background climate. In this context, in addition to lowering the annual mean insolation by obliquity, reducing meridional heat transport by the AMOC should also be able to cool down the northern high latitudes further during Heinrich Stadials. This is corroborated by an extreme LGM North Atlantic hosing experiment (with 0.2 Sv freshwater input) in which the AMOC shuts down (LGM_NA02) (Extended Data Table 1 and Extended Data Fig. 5d–g). As expected, an overall cooling appears in the subsurface of the northern North Pacific, although the simulated Aleutian Low and tropical warming get even stronger.

CIS meltwater events and North Atlantic circulation. To qualify the impact of CIS meltwater events on North Atlantic circulation during Heinrich Stadial 1, we performed a North Pacific-alone hosing experiment (0.1 Sv) under LGM boundary conditions (LGM_NP) (Extended Data Table 1). The experiment was integrated for 600 model years, and the average of the last 100 years is used to represent the corresponding climatology. It appears that North Pacific-alone hosing leads to robust subsurface warming in the North Pacific (Extended Data Fig. 5b); this warming acts as a positive feedback to maintain and/or accelerate the retreat of marine-based CIS ice. Note that, for the North Pacific, the term ‘subsurface’ is used for water depths of 120–180 m, according to the glacial northeastern North Pacific shelf depth. The North Atlantic, on the other hand, is characterized by discernible subsurface cooling (below around 100 m) (Extended Data Fig. 5c). Assuming that Heinrich events are related to subsurface warming (at roughly 100–1,200 m) in the North Atlantic^{28,71,72}, subsurface cooling would rather hamper the occurrences of Heinrich events by stabilizing the marine-based ice in the North Atlantic region. Therefore, it is more likely that changes in the North Atlantic are triggering North Pacific ice-raftering events during Heinrich Stadials than vice versa.

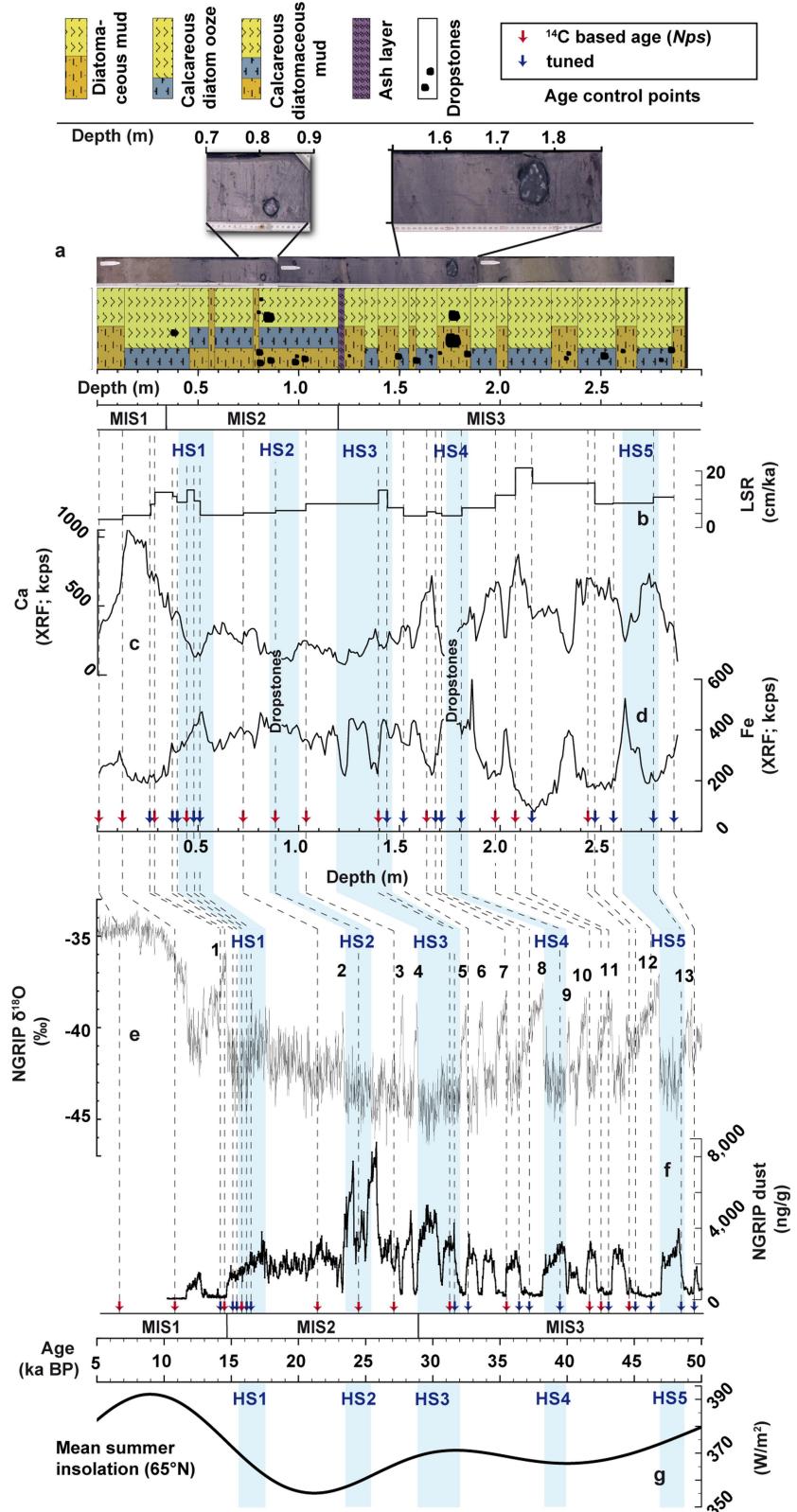
Code availability. The standard model code of the Community Earth System Models (COSMOS) version COSMOS-landveg r2413 (2009) is available upon request from the Max Planck Institute for Meteorology in Hamburg (<https://www.mpimet.mpg.de>). The code for the BACON software used for age-model construction can be obtained from <http://www.chrono.qub.ac.uk/blaauw/>.

Data availability. Our data can be obtained from the PANGAEA database at <https://pangaea.de> (<https://doi.org/10.1594/PANGAEA.887506>) and/or can be found in the Extended Data. No statistical methods were used to predetermine sample size.

31. Ruth, U. et al. Ice core evidence for a very tight link between North Atlantic and east Asian glacial climate. *Geophys. Res. Lett.* **34**, L03706 (2007).
32. Serno, S. et al. Comparing dust flux records from the Subarctic North Pacific and Greenland: implications for atmospheric transport to Greenland and for application of dust as a chronostratigraphic tool. *Paleoceanography* **30**, 583–600 (2015).
33. Hundy, I. L., Kennett, J. P., Roark, E. B. & Ingram, B. L. Apparent synchronicity of submillennial scale climate events between Greenland and Santa Barbara Basin, California from 30–10 ka. *Quat. Sci. Rev.* **21**, 1167–1184 (2002).
34. Cartapanis, O., Tachikawa, K. & Bard, E. Northeastern Pacific oxygen minimum zone variability over the past 70 kyr: impact of biological production and oceanic ventilation. *Paleoceanography* **26**, PA4208 (2011).
35. Blaauw, M. & Christen, A. J. Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Anal.* **6**, 457–474 (2011).
36. Reimer, P. J. et al. INTCAL13 and MARINE13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2013).
37. Sarnthein, M., Balmer, S., Grootes, P. M. & Muddelsee, M. Planktic and benthic ^{14}C reservoir ages for three ocean basins calibrated by a suite of ^{14}C plateaus in the glacial-to-deglacial Suigetsu atmospheric ^{14}C record. *Radiocarbon* **57**, 129–151 (2015).
38. Maier, E. et al. Combined oxygen and silicon isotope analysis of diatom silica from a deglacial subarctic Pacific record. *J. Quat. Sci.* **28**, 571–581 (2013).
39. Sancetta, C. Distribution of diatom species in surface sediments of the Bering and Okhotsk seas. *Micropaleontology* **28**, 221–257 (1982).
40. Sancetta, C. Three species of *Coscinodiscus* Ehrenberg from North Pacific sediments examined in the light and scanning electron microscopes. *Micropaleontology* **33**, 230–241 (1987).
41. Swann, G. E. A. & Leng, M. J. A review of diatom $\delta^{18}\text{O}$ in paleoceanography. *Quat. Sci. Rev.* **28**, 384–398 (2009).
42. Taylor, H. P. J. The oxygen isotope geochemistry of igneous rocks. *Contrib. Mineral. Petro.* **19**, 1–71 (1968).
43. Sheppard, A. M. F. & Gilg, H. A. Stable isotope geochemistry of clay minerals. *Clay Miner.* **31**, 1–24 (1996).
44. Chaplgin, B. et al. A high-performance, safer and semi-automated approach for the analysis of diatom silica and new methods for removing exchangeable oxygen. *Rapid Commun. Mass Spectrom.* **24**, 2655–2664 (2010).
45. Chaplgin, B. et al. Inter-laboratory comparison of oxygen isotope compositions from biogenic silica. *Geochim. Cosmochim. Acta* **75**, 7242–7256 (2011).
46. Von Huene, R., Larson, E. & Crouch, J. in *Initial Reports of the Deep Sea Drilling Project*, Vol. XVIII (eds Misch, L. F. & Weser, O. E.) 835–842 (US Gov. Printing Office, Washington DC, 1973).
47. St John, K. E. K. & Krissek, L. A. Regional patterns of Pleistocene ice rafted debris flux in the North Pacific. *Paleoceanography* **14**, 653–662 (1999).
48. Méheust, M., Fahl, K. & Stein, R. Variability in modern sea surface temperature, sea ice and terrigenous input in the sub-polar North Pacific and Bering Sea: reconstruction from biomarker data. *Org. Geochem.* **57**, 54–64 (2013).
49. Sikes, E. L., Volkman, J. K., Robertson, L. G. & Pichon, J.-J. Alkenones and alkenes in surface waters and sediments of the Southern Ocean: implications for paleotemperature estimation in polar regions. *Geochim. Cosmochim. Acta* **61**, 1495–1505 (1997).
50. Leclerc, A. J. & Labeyrie, L. Temperature dependence of the oxygen isotopic fractionation between diatom silica and water. *Earth Planet. Sci. Lett.* **84**, 69–74 (1987).
51. Waelbroeck, C. et al. Sea-level and deep water temperature changes derived from benthic foraminifera isotopic records. *Quat. Sci. Rev.* **21**, 295–305 (2002).
52. Takahashi, K. Seasonal fluxes of pelagic diatoms in the subarctic Pacific, 1982–1983. *Deep-Sea Res.* **33**, 1225–1251 (1986).
53. Epstein, S. & Sharp, R. P. Oxygen-isotope variations in the Malaspina and Saskatchewan glaciers. *J. Geol.* **67**, 88–102 (1959).
54. Dansgaard, W. & Tauber, H. Glacier oxygen-18 content and Pleistocene ocean temperatures. *Science* **166**, 499–502 (1969).
55. IAEA/WMO. Global Network of Isotopes in Precipitation. *The GNIP Database* <http://www.iaea.org/water> (2015).
56. Jakobsson, M. et al. Post-glacial flooding of the Bering Land Bridge dated to 11 calka BP based on new geophysical and sediment records. *Clim. Past* **13**, 991–1005 (2017).
57. Roeckner, E. et al. *The Atmospheric General Circulation Model ECHAM5. Part 1: Model Description* (Max Planck Inst. Meteorol. Rep. 349, 2003).
58. Brovkin, V., Raddatz, T., Reick, C. H., Claussen, M. & Gaylor, V. Global biogeophysical interactions between forest and climate. *Geophys. Res. Lett.* **36**, L07405 (2009).
59. Marsland, S. J., Haak, H., Jungclaus, J. H., Latif, M. & Röske, F. The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. *Ocean Model.* **5**, 91–127 (2003).
60. Hibler, W. III A dynamic thermodynamic sea ice model. *J. Phys. Oceanogr.* **9**, 815–846 (1979).
61. Zhang, X., Lohmann, G., Knorr, G. & Purcell, C. Abrupt glacial climate shifts controlled by ice sheet changes. *Nature* **512**, 290–294 (2014).
62. Zhang, X., Knorr, G., Lohmann, G. & Barker, S. Abrupt North Atlantic circulation changes in response to gradual CO_2 forcing in a glacial climate state. *Nat. Geosci.* **10**, 518–523 (2017).
63. Monterey, G. & Levitus, S. *Seasonal Variability of Mixed Layer Depth for the World Ocean* (NOAA Atlas NESDIS 14, US Gov. Printing Office, Washington DC, 1997).
64. Kuroyanagi, A., Kawahata, H. & Nishi, H. Seasonal variation in the oxygen isotopic composition of different-sized planktonic foraminifer *Neogloboquadrina pachyderma* (sinistral) in the northwestern North Pacific and implications for reconstruction of the paleoenvironment. *Paleoceanography* **26**, PA4215 (2011).
65. Mulitza, S. et al. Temperature: $\delta^{18}\text{O}$ relationships of planktonic foraminifera collected from surface waters. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **202**, 143–152 (2003).
66. Zhang, Y. et al. Equatorial Pacific forcing of western Amazonian precipitation during Heinrich Stadial 1. *Sci. Rep.* **6**, 35866 (2016).
67. Lambek, K. & Chappell, J. Sea level change through the last glacial cycle. *Science* **292**, 679–686 (2001).
68. Grant, K. M. et al. Rapid cooling between ice volume and polar temperature over the past 150,000 years. *Nature* **491**, 744–747 (2012).
69. Köhler, P., Nehrbass-Ahles, C., Schmitt, J., Stocker, T. F. & Fischer, H. A. 156 kyr smoothed history of the atmospheric greenhouse gases CO_2 , CH_4 , and N_2O and their radiative forcing. *Earth Syst. Sci. Data* **9**, 363–387 (2017).
70. Berger, A. L. Long-term variations of caloric insolation resulting from the Earth's orbital elements. *Quat. Res.* **9**, 139–167 (1978).
71. Flückinger, J., Knutti, R. & White, J. W. C. Oceanic processes as potential trigger and amplifying mechanisms for Heinrich events. *Paleoceanography* **21**, PA2014 (2006).
72. Alvarez-Solas, J. et al. Heinrich event 1: an example of dynamic ice-sheet reaction to oceanic changes. *Clim. Past* **7**, 1297–1306 (2011).
73. Bronk Ramsey, C. et al. A complete terrestrial radiocarbon record for 11.2 to 52.8 kyr B.P. *Science* **338**, 370–374 (2012).
74. Gersonde, R. Documentation of sediment core SO202-27-6 (2010).
75. Laskar, J. et al. A long-term numerical solution for the insolation quantities of the Earth. *Astron. Astrophys.* **428**, 261–285 (2004).
76. Clague, J. J. & James, T. S. History and isostatic effects of the last ice sheet in southern British Columbia. *Quat. Sci. Rev.* **21**, 71–87 (2002).
77. Kaufman, D. S., Young, N. E., Briner, J. P. & Manley, W. F. Alaska palaeo-glacier atlas (version 2). *Dev. Quat. Sci.* **15**, 427–445 (2011).
78. Ehlers, J. & Gibbard, P. L. The extent and chronology of Cenozoic global glaciation. *Quat. Int.* **164–165**, 6–20 (2007).
79. Schlitzer, R. *Ocean Data View* <https://odv.awi.de> (2018).

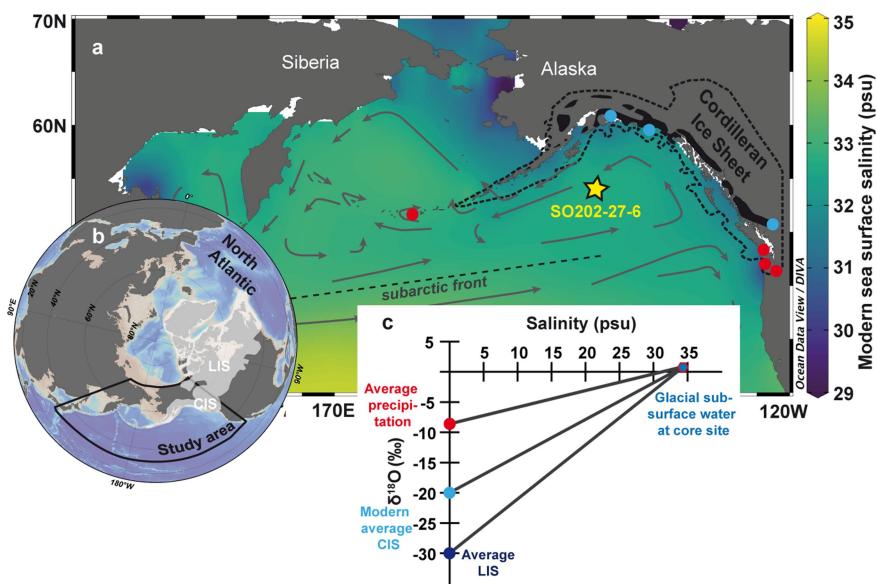


Extended Data Fig. 1 | Results from two freshwater hosing experiments. Left, LGM_NA; right LGM_NA+NP. Model results are presented as anomalies between the hosing simulations and the LGM state (see Methods). **a, e**, Total precipitation anomalies. **b, f**, Precipitation $\delta^{18}\text{O}$ anomalies. **c, g**, $\delta^{18}\text{O}_{\text{diat.}}$ anomalies. **d, h**, Subsurface $\delta^{18}\text{O}_{\text{sw}}$ anomalies (at depths of 120–180 m). The yellow star marks the location of core SO202-27-6.



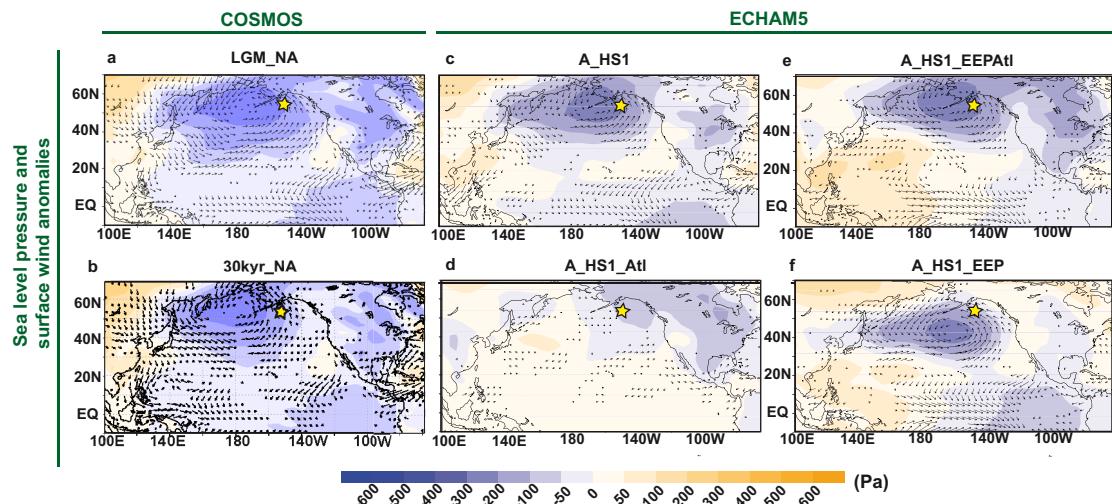
Extended Data Fig. 2 | Link between our data from core SO202-27-6, and NGRIP climate variabilities. a-d, SO202-27-6. a, Scheme and pictures describing the sediment core⁷⁴. b, Linear sedimentation rate (LSR). c, Calcium intensity based on XRF analysis. d, Iron intensity based

on XRF analysis. ka, thousands of years ago; kcps, thousands of counts per second. e, NGRIP $\delta^{18}\text{O}$ record⁷. f, NGRIP dust concentration³¹, including age-control points for SO202-27-6. g, Mean summer insolation at 65° N (ref. ⁷⁵).



Extended Data Fig. 3 | SSS/δ¹⁸O_{sw} mixing model for the last glacial open northeastern North Pacific. a, Study area as shown in Fig. 1, including the modern extent of the CIS (black), the extent of the CIS during the LGM^{76,77} (black dashed line), the site of the studied core (yellow star), and the locations where precipitation (red dots) and modern glacier (light blue dots) parameters were taken for the SSS/δ¹⁸O_{sw} mixing model. **b**, Study

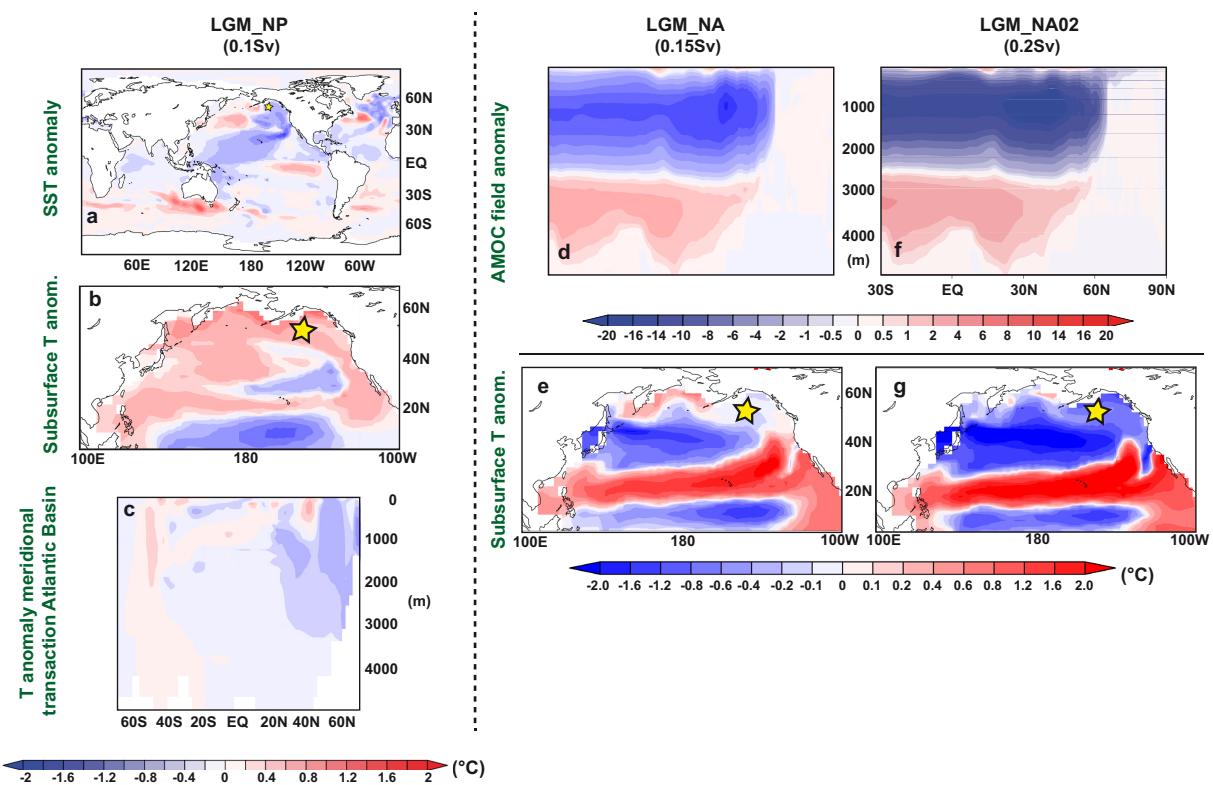
area from the North Pacific. Shaded white areas represent the extents of the LIS and CIS during the LGM⁷⁸. **c**, SSS/δ¹⁸O_{sw} mixing model assuming linear regression between SSS and δ¹⁸O_{sw}. We used three low-salinity endmembers and one high-salinity endmember to estimate SSS changes at our core site between 43 kyr and 11 kyr ago (see Methods). The Northern Hemisphere map and the SSS map were created using Ocean Data View⁷⁹.



Extended Data Fig. 4 | Sea-level-pressure and surface-wind anomalies in our hosing experiments. a, b, Results obtained using COSMOS.

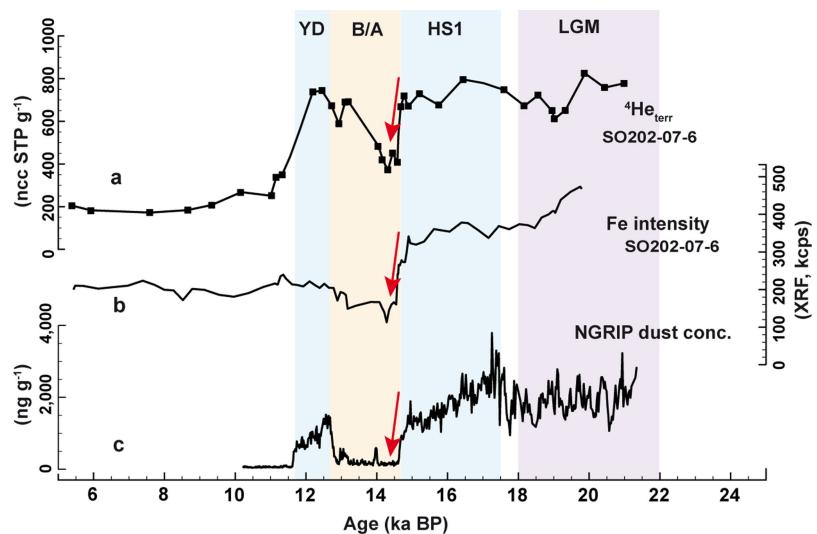
a, LGM_NA experiment. b, 30kyr_NA experiment. c-f, Results obtained using ECHAM5. c, A_HS1 experiment. d-f, A_HS1 experiment, imposing SST fields on the Atlantic Basin (Atl) only (d), the Atlantic Basin and the

east Equatorial Pacific (EEP; e), and the EEP only (f). The yellow star marks the location of studied Core SO202-27-6. Surface-wind anomalies (vectors) are presented in m s^{-1} . Sea-level-pressure anomalies are shown with shading.



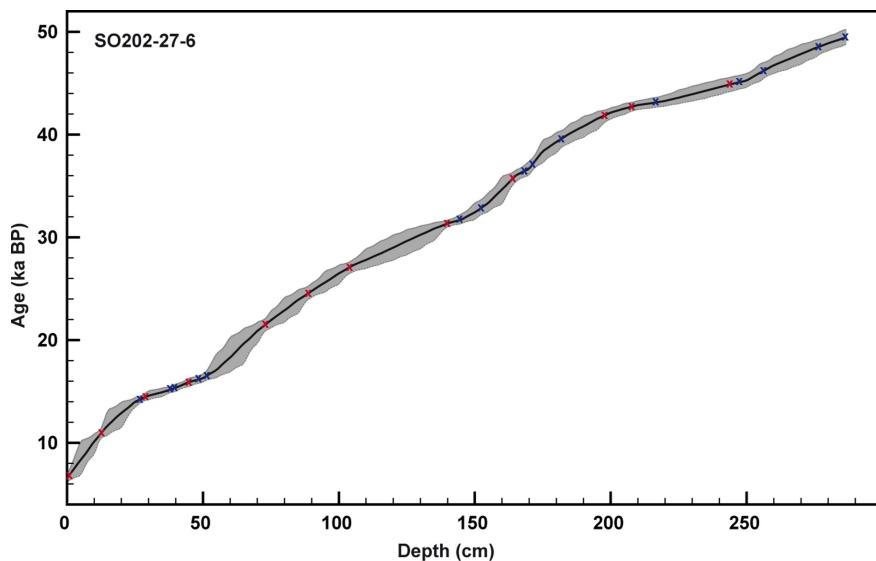
Extended Data Fig. 5 | Results from freshwater hosing experiments LGM_NP, LGM_NA and LGM_NA02, presented as anomalies.
a–c, LGM_NP experiment (0.1 Sv). **a**, Global SST anomaly; **b**, North Pacific subsurface temperature anomaly (120–180 m). **c**, Temperature anomaly over the meridional transaction of the Atlantic basin (60° W

to 15° W). **d–g**, LGM_NA experiment (0.15 Sv) (**d, e**) and LGM_NA02 experiment (0.2 Sv) (**f, g**). **d, f**, AMOC field anomalies. **e, g**, Subsurface temperature anomalies (120–180 m). The yellow star marks the location of core SO202-27-6.



Extended Data Fig. 6 | Comparison of northwestern North Pacific eolian dust and iron intensity, as well as NGRIP dust concentration over the last deglaciation. **a**, Eolian dust (terrestrial ${}^4\text{He}$ concentration) 32 and **b**, iron intensity from core SO202-07-6 (51.3° N, 167.7° E; 2,340 m water depth). **c**, NGRIP dust concentration 31 . Dust changes in the northwestern North Pacific and Greenland are synchronous 32 , and coincide with

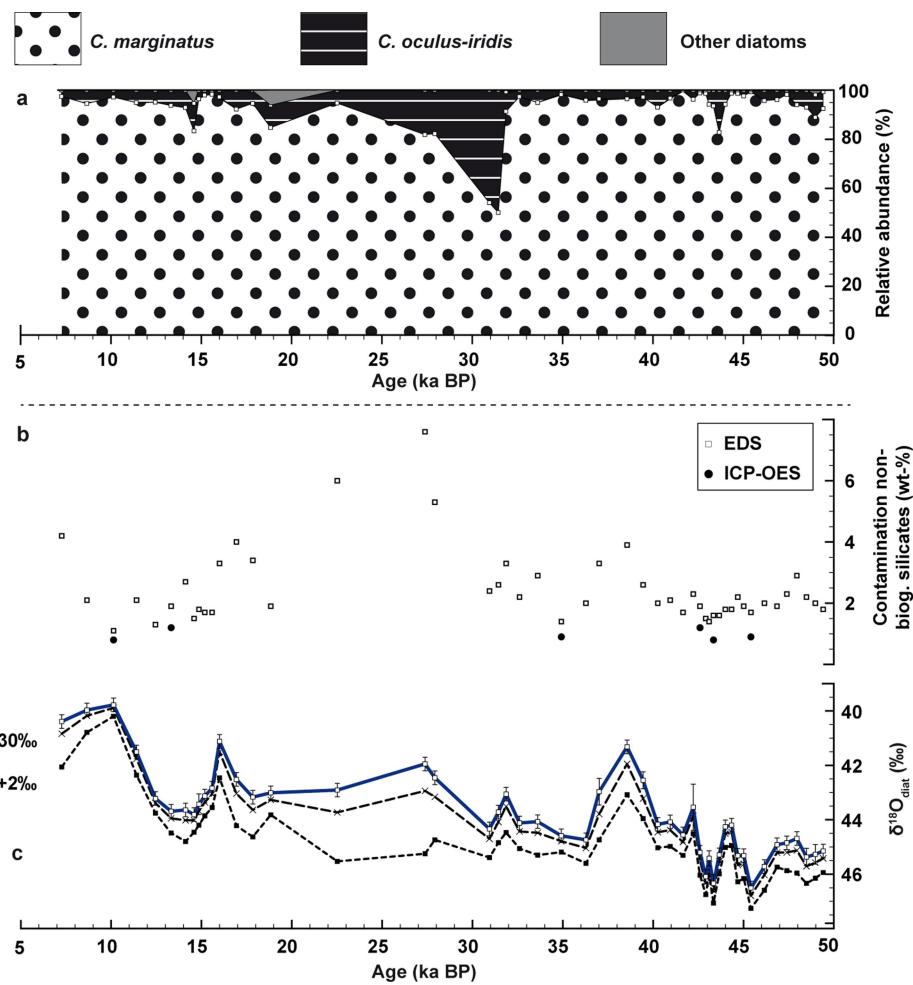
iron-intensity changes in the northwestern North Pacific. B/A, Bølling/Allerød interstadial; YD, Younger Dryas cold period. Red arrows mark chronological coincidence between the changes in ${}^4\text{He}$, iron intensity and NGRIP dust concentration; ncc STP g $^{-1}$, nano-cubic centimetre per gram at standard temperature and pressure.



Extended Data Fig. 7 | Age–depth relationship for core SO202-27-6.

The grey envelope shows the 95% confidence interval around the median age (black line). Crosses indicate age-control points obtained from

radiocarbon dating (red), and from proxy correlation to core MD02-2489 (ref. ¹²) and the NGRIP dust record³¹ (blue).



Extended Data Fig. 8 | Diatom isotope sample composition, residual contamination with non-biogenic silicates and mass-balance-corrected $\delta^{18}\text{O}_{\text{diat}}$. (from core SO202-27-6. a, Relative abundances of the following diatom species in the isotope samples: *C. marginatus*, *C. oculus-iridis*, and other diatom species. b, Contamination of purified diatom samples with non-biogenic silicates, estimated by inductively coupled plasma optical emission spectrometry (ICP-OES) and energy-dispersive X-ray spectrometry (EDS). c, Blue line, measured $\delta^{18}\text{O}_{\text{diat}}$ values (error

bars indicate errors of replicate analyses or long-term reproducibility of standards (1σ)). Black dotted lines, $\delta^{18}\text{O}_{\text{diat}}$ values that have been mass-balance-corrected for contamination with non-biogenic silicates (estimated by EDS), and using one of two different $\delta^{18}\text{O}$ values for non-biogenic silicate contamination ($+2\text{\textperthousand}$ or $+30\text{\textperthousand}$). Contamination values, $\delta^{18}\text{O}_{\text{diat}}$ values and mass-balance-corrected $\delta^{18}\text{O}_{\text{diat}}$ values younger than 25 kyr BP are taken from ref. ¹².

Extended Data Table 1 | Overview of model experiments

a COSMOS	Boundary conditions	Orbital parameters	FWF amount	FWF	Integrated
				IsoV	years
LGMctl	21 ka	21 ka	-	-	5000
LGM_NA	21 ka	21 ka	0.15 Sv NA	30‰	800
LGM_NA+NP	21 ka	21 ka	0.15 Sv NA + 0.1 Sv NP	30‰	800
LGM_NA02	21 ka	21 ka	0.2 Sv NA	30‰	800
LGM_NP	21 ka	21 ka	0.1 Sv NP	30‰	600
30kyr_ctl	21 ka	30 ka	-	-	800
30kyr_NA	21 ka	30 ka	0.15 Sv NA	30‰	600

b ECHAM5	Boundary conditions	SST forcing
A_LGM	21 ka	LGMctl global SST
A_HS1	21 ka	LGM_NA global SST
A_HS1_Atl	21 ka	LGMctl SST background, but with LGM_NA SST only in the Atlantic
A_HS1_EEPAtl	21 ka	LGMctl SST background, but with LGM_NA SST in the Atlantic and EEP
A_HS1_EEP	21 ka	LGMctl SST background, but with LGM_NA SST only in the EEP

a, Experiments conducted with the fully coupled GCM COSMOS. The boundary conditions include ice-sheet configuration, land-sea mask (land mask that is related to the sea-level changes in the past), glacial extent, greenhouse gases, etc., except the orbital parameters (see Methods). FWF, freshwater forcing; IsoV, isotopic values; NA, North Atlantic (40–55° N, 20–45° W); NP, North Pacific (50–60° N, 143–172° W); 1 Sv = $10^6 \text{ m}^3 \text{ s}^{-1}$. ‘Years’ here refers to model years.

b, Experiments conducted with ECHAM5 (the atmospheric component of COSMOS).

Extended Data Table 2 | Age constraints of core SO202-27-6

Sample ID/ Age control point	Depth (cm)	¹⁴ C ages (ka)	¹⁴ C age error (a)	Reservoir age (Delta-R) (a)	Reservoir age error (1 σ) (a)	Lake Suigetsu varve error (1 σ) (a)	NGRIP tuning error (1 σ) (a)	Median ages (ka BP)
OS-85661	0.5	6.090	30	40	173	-	-	6.713
OS-85752	12.5	9.880	30	40	173	-	-	10.923
LS-1	26.5	-	-	-	-	102	-	14.173
OS-87903	28.5	13.050	55	40	173	108	-	14.415
LS-2	37.5	-	-	-	-	132	-	15.137
LS-3	39.5	-	-	-	-	148	-	15.318
OS-85753	44.5	13.900	30	150	185	165	-	15.876
LS-4	48.0	-	-	-	-	179	-	16.140
LS-5	51.0	-	-	-	-	190	-	16.459
OS-87888	72.5	18.750	70	710	202	-	-	21.401
OS-87894	88.5	21.400	120	710	202	-	-	24.500
OS-87892	103.5	23.800	110	710	202	-	-	27.022
OS-88043	139.5	28.300	140	710	202	-	-	31.303
NG-1	144.0	-	-	-	-	-	375	31.643
NG-2	152.0	-	-	-	-	-	375	32.792
OS-87893	163.5	33.000	170	710	202	-	-	35.594
NG-3	168.0	-	-	-	-	-	375	36.410
NG-4	171.0	-	-	-	-	-	375	37.013
NG-5	181.0	-	-	-	-	-	375	39.441
OS-87899	197.5	38.400	310	710	202	-	-	41.823
OS-87889	207.5	39.900	290	710	202	-	-	42.694
NG-6	216.0	-	-	-	-	-	375	43.096
OS-87898	243.5	42.800	600	710	202	-	-	44.854
NG-7	247.0	-	-	-	-	-	375	45.076
NG-8	256.0	-	-	-	-	-	375	46.158
NG-9	276.0	-	-	-	-	-	375	48.485
NG-10	286.0	-	-	-	-	-	375	49.413

Apart from planktic (sinistral *N. pachyderma*) ¹⁴C ages, additional age-control points were obtained through correlation to the high-resolution Lake Suigetsu record—via proxy correlation to core MD02-2489 (ref. ¹²)—and through proxy correlation to the NGRIP dust record (see Methods). Radiocarbon ages and radiocarbon-age errors from the upper 90 cm are taken from ref. ¹². Reservoir ages and reservoir-age errors were assigned from nearby core MD02-2489 (ref. ³⁷). Lake Suigetsu varve errors are taken from ref. ⁷³, a, years ago.

Evolution of cooperation in stochastic games

Christian Hilbe^{1,2*}, Štěpán Šimsa³, Krishnendu Chatterjee^{2*} & Martin A. Nowak^{1,4*}

Social dilemmas occur when incentives for individuals are misaligned with group interests^{1–7}. According to the ‘tragedy of the commons’, these misalignments can lead to overexploitation and collapse of public resources. The resulting behaviours can be analysed with the tools of game theory⁸. The theory of direct reciprocity^{9–15} suggests that repeated interactions can alleviate such dilemmas, but previous work has assumed that the public resource remains constant over time. Here we introduce the idea that the public resource is instead changeable and depends on the strategic choices of individuals. An intuitive scenario is that cooperation increases the public resource, whereas defection decreases it. Thus, cooperation allows the possibility of playing a more valuable game with higher payoffs, whereas defection leads to a less valuable game. We analyse this idea using the theory of stochastic games^{16–19} and evolutionary game theory. We find that the dependence of the public resource on previous interactions can greatly enhance the propensity for cooperation. For these results, the interaction between reciprocity and payoff feedback is crucial: neither repeated interactions in a constant environment nor single interactions in a changing environment yield similar cooperation rates. Our framework shows which feedbacks between exploitation and environment—either naturally occurring or designed—help to overcome social dilemmas.

The tragedy of the commons leads to the question of how to manage and conserve public resources^{1–8}. Any solution to this problem requires an understanding of which processes drive human cooperation and how institutions, norms and other feedback mechanisms can be used to reinforce positive behaviours²⁰. These questions are often explored by analysing stylized social dilemmas, such as the public goods game²¹ or the collective-risk dilemma²², that provide valuable insights into the dynamics of cooperation in controlled settings. When subjects interact in such games over multiple rounds, it is typically assumed that the public good remains constant in time, independent of the outcome of previous interactions^{9–15}. Here, we explore the emergence of reciprocity when strategic choices in one round affect game payoffs in subsequent rounds. We introduce a framework that allows us to capture the idea that humans affect and are affected by the value of the public resource, and that they are able to anticipate and to adapt to such endogenous changes.

Our approach is based on the theory of stochastic games^{16,17}. A group of players can find itself in one of multiple states (Fig. 1). The different states capture how the present physical or social environment affects the feasible actions of the players and their payoffs. The theory of stochastic games^{16–19} has applications in computer science^{23,24}, industrial organization, capital accumulation and resource extraction¹⁷.

We consider stochastic games where, in each state, players interact in a social dilemma with different payoff values. The decision by the players of whether to cooperate or to defect not only affects their current payoffs but also the game that will be played in the next round. In Fig. 1 we illustrate a scenario that reflects the tragedy of the commons. Mutual cooperation improves the quality of the public resource, leading the players to interact in game 1 with comparably high payoffs. Partial defection leads to a deterioration of the resource; players move to game 2 where payoffs are lower. The stochastic game is played for

many rounds. Transitions between different states can be stochastic or deterministic, state-dependent or state-independent. The well-studied framework of repeated games is a special case of stochastic games with only one state.

The effect of changing environments on evolutionary dynamics has been explored previously in one-shot, non-repeated games, not using the theory of stochastic games^{25–29} (see Supplementary Information, section 1.1). In some scenarios, the co-evolution of the players’ strategies and their environment can lead to oscillations between cooperators and defectors^{27,28}. But if cooperators are at a disadvantage in every environment, environmental feedback is ineffective to prevent cooperators from going extinct (Supplementary Information). One-shot models assume that players consider only their present payoff when making strategic choices. In stochastic games, players take a long-term perspective instead. To find optimal strategies, they need to consider how their actions affect the response of their opponents and the future state of the environment. As we show, this interplay between reciprocity and payoff feedback can be crucial for cooperation.

Traditionally, work on stochastic games considers rational players who can employ arbitrarily complex strategies, but does not focus on the dynamics of how players adapt their strategies. We introduce an evolutionary perspective to stochastic games. Players do not need to act rationally, but instead they experiment with available strategies and imitate others depending on success³⁰. We use simple strategies that are easy to implement and to interpret⁸. Such an evolutionary set-up has proved useful to understand the dynamics of cooperation in repeated games^{8–13}.

We first study a stochastic game with two states (Fig. 2). Individuals use pure ‘memory one’ strategies whereby a player’s move depends on only the present state and the outcome of the previous round (see Methods and Supplementary Information for details). We compare the stochastic game with the two associated repeated games where the same game occurs every round (Fig. 2). We consider two-player interactions that represent prisoner’s dilemmas, as well as n -player public-goods games. In both cases, cooperation entails a cost $c > 0$. In the prisoner’s dilemma, cooperation yields a benefit $b_i > c$ to the co-player, where b_i depends on the state i . In the public goods game, aggregated costs are multiplied by a factor r_i (with $1 < r_i < n$ depending on state i), and redistributed among all players. Game 1 is more profitable than game 2 if $b_1 > b_2$ or $r_1 > r_2$. Players find themselves in game 1 only if everyone has cooperated in the previous round. Our simulations show that this feedback can boost cooperation markedly. For reasonable parameters, the stochastic game populations adapt quickly towards full cooperation, although neither of the two repeated games alone yields substantial cooperation levels.

In the stochastic game, cooperation evolves because defectors lose out twice: once, because they risk receiving less cooperation from reciprocal co-players in future and second, because players collectively move towards a less beneficial game. The stochastic game is most effective in boosting cooperation if the benefit in game 1 is intermediate (Extended Data Fig. 1). If b_1 is too low, the double loss present in the stochastic game is not sufficient to incentivize mutual cooperation, whereas if b_1 is high, players cooperate in the first game anyway. Stochastic games can lead to cooperation even if all individual repeated games fail.

¹Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA. ²IST Austria, Klosterneuburg, Austria. ³Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic. ⁴Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, MA, USA. *e-mail: christian.hilbe@ist.ac.at; krishnendu.chatterjee@ist.ac.at; martin_nowak@harvard.edu

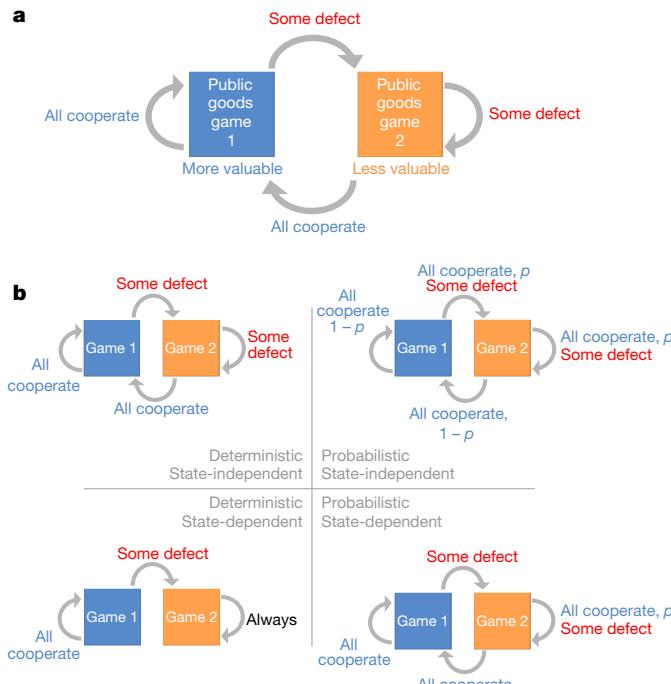


Fig. 1 | In stochastic games, the decisions made by players in one round determine the game that will be played next round. **a**, For example, if some players defect in a public-goods game, then the environment could deteriorate and thereby reduce the value of the public good. If all cooperate, then the environment could recover and the original value of the public good might be restored. The different states of the environment correspond to the different games that can be played. In this illustration, we show two public-goods games with $r_1 > r_2$. **b**, A stochastic game is deterministic if the players' actions and the current game uniquely determine the game that will be played next round. It is state-independent if the game in the next round depends on only the players' actions, not the current game (state). Thus, we distinguish four different types of stochastic game, depending on whether transitions are deterministic or probabilistic (where p and $1 - p$ indicate the probability of making the respective transition), and whether they are state-independent or state-dependent. We note that even a game that involves only deterministic transitions is referred to as a 'stochastic' game, because it represents a special case of the framework.

We derive a condition for the stability of cooperation in stochastic games with two states and state-independent transitions. A numerical analysis for the two-player case suggests that full cooperation emerges when win-stay lose-shift⁹ (WSLS) becomes stable (Extended Data Figs. 2, 3). This strategy prescribes cooperation in the next round if and only if both players used the same action in the previous round. In a conventional repeated prisoner's dilemma, WSLS is a Nash equilibrium if $b \geq 2c$ (ref. ⁸). In the stochastic game, WSLS is an equilibrium if

$$(2q_2 - q_0)b_1 + (1 - 2q_2 + q_0)b_2 \geq 2c \quad (1)$$

where the parameters q_i refer to the conditional probability that the players will be in game 1 in the next round given that i of them have cooperated in the present round. If mutual cooperation leads to game 1 and mutual defection to game 2, then $q_2 = 1$ and $q_0 = 0$. Therefore, WSLS is stable if $2b_1 - b_2 \geq 2c$. Because $b_1 > b_2$, this condition is easier to satisfy than the respective conditions for the two associated repeated games.

The condition in equation (1) highlights the fact that the stability of cooperation depends on how the states change given the players' decisions. To explore the effect of this exogenous feedback systematically, we perform simulations for all eight deterministic and state-independent two-state games (Extended Data Fig. 2). In six of the eight cases, players spend more time in the profitable game 1. But only in

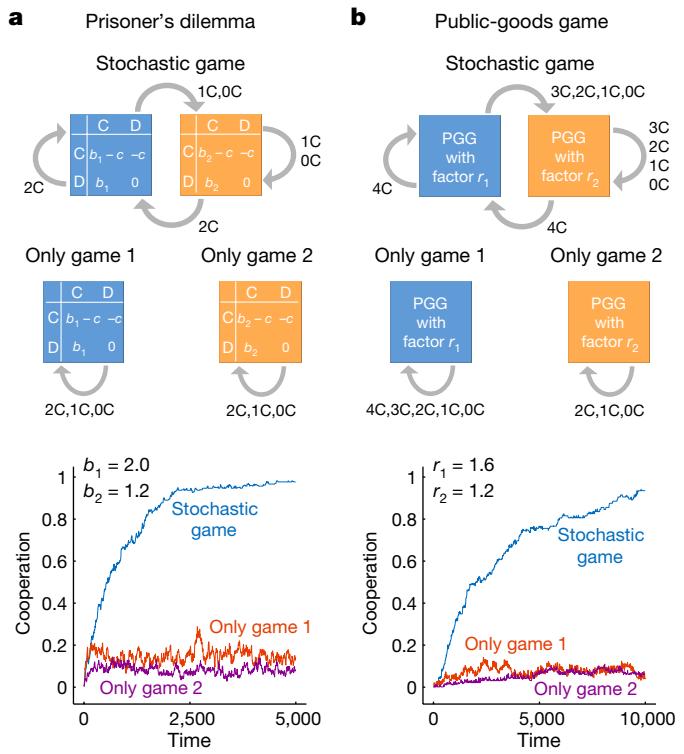


Fig. 2 | Stochastic games can promote cooperation even if all individual games favour defection. **a, b**, We study the repeated prisoner's dilemma, which is a two-player game (**a**), and the repeated public-goods game (PGG), which is interpreted here as a four-player game (**b**). In both cases, the first game has a higher benefit from cooperation than the second game. Arrows represent the possible transitions, and the arrow labels indicate the number of co-operators ('C') required for the respective transition. The two-player games are represented by their payoff matrices. In the stochastic game, if all players cooperate then the next round will be the first game, but if some players defect ('D') then the next round will be the second game. In the standard repeated games, the same game is used in every round. An analysis based on evolutionary dynamics reveals that each of the standard repeated games fails to support cooperation, whereas the stochastic game favours cooperation. The time axis corresponds to the number of mutant strategies introduced into the population (see Methods). Parameter values: **a**, $b_1 = 2$, $b_2 = 1.2$, $c = 1$; **b**, $r_1 = 1.6$, $r_2 = 1.2$, $c = 1$.

two of them do players actually cooperate. In line with equation (1), cooperation evolves only if $q_2 = 1$ and $q_0 = 0$, with q_1 being irrelevant. Stochastic games are most effective in promoting cooperation if mutual cooperation improves the public good while mutual defection deteriorates it—a natural scenario. Analogous conclusions hold for multiplayer interactions (Extended Data Figs. 4, 5).

Probabilistic transitions can further enhance the evolution of cooperation. In Fig. 3a, mutual cooperation in game 2 leads back to game 1 with probability q . The optimal value of q is intermediate: players should have some chance to return to the better state, but it should not be too easy (see also Extended Data Fig. 6). In Fig. 3b, the length of the game is not exogenously given, but affected by the players' decisions. Individuals start in state 1, in which they play a conventional prisoner's dilemma; if one or both players defect, then there is some probability q that players move towards state 2, in which no further profitable interactions are possible. This form of environmental feedback promotes cooperation; payoffs become maximal for small but positive q (Extended Data Fig. 7). In Fig. 3c we consider a model with timeout. Defection leads to a temporal state in which no profitable interactions are possible. The return probability to the regular game is q . We derive adaptive dynamics for simple reactive strategies (x, y) , where x denotes the cooperation probability after having been in state 1 previously and y is the cooperation probability after having been in timeout. We find

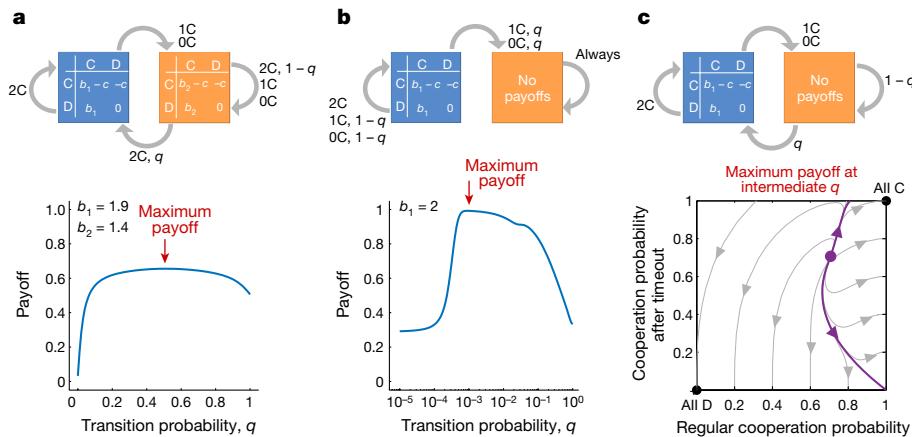


Fig. 3 | Probabilistic transitions maximize cooperation in three different stochastic games. **a**, Game 1 is more profitable than game 2, but mutual cooperation in game 2 leads to game 1 only with probability q . The evolving average payoffs are maximized for intermediate q . **b**, Game 1 is left with probability q if at least one player has defected. The optimal value of q is small but positive for games with a finite number of rounds (continuation probability $\delta = 0.999$). **c**, Defection leads to a timeout with an expected duration that depends on the return probability q . We derive

that the fully cooperative strategy $(1, 1)$ can become stable, although unconditional cooperation is never stable in a conventional repeated prisoner's dilemma.

Next we explore the ideal feedback between game payoff and strategic choice. We consider a stochastic game with four players and five states. Defection by a subgroup of players has an immediate, gradual or delayed negative impact on the benefits of cooperation, or no effect (Fig. 4). We obtain the highest cooperation rates for immediate negative impact. The intuitive explanation is as follows: maximum cooperation arises if the players are most incentivized to cooperate in the most valuable game. In the immediate scenario, any deviation from cooperation in

the adaptive dynamics for strategies that take into account only whether players have been in game 1 in the previous round or in the timeout. Depending on the parameters, 'All C' is a stable endpoint of evolution because no nearby mutant strategy can yield a higher payoff. Again the optimal value of q is intermediate: low values of q increase the area of the phase space for which populations move towards cooperation, but they also make occasional errors more costly (parameters $b_1 = 3$, $c = 1$, $q = 1/2$).

game 1 leads to a game with the lowest payoff. Interestingly, even the scenario with a delayed response promotes higher cooperation rates than the game in which the public good remains unchanged across all states. The lowest cooperation rates are obtained when the benefits of cooperation are high in all five games. We obtain similar conclusions for a state-dependent game in which it takes several successive rounds of mutual defection to end up in the worst state (Extended Data Figs. 8, 9).

Direct reciprocity is a mechanism for the evolution of cooperation based on repeated interactions. The standard assumption has been that the same game, with the same payoff, is played again and again. We have introduced the concept that the game payoff changes in different rounds. We explore cases in which cooperation leads to a more valuable game next round and defection to a less valuable one. Surprisingly, we find that this setting boosts cooperation markedly. In the resulting stochastic game, cooperation can prevail even if it is unsuccessful in all individual repeated games. Our observations suggest how naturally occurring or designed feedback can promote cooperation. A tragedy of the commons can be avoided if the environment deteriorates (rapidly) as a consequence of defection. Likewise, cooperation is boosted if there is the prospect of playing for higher gains should the current cooperation succeed. The evolutionary analysis of stochastic games represents a new tool for understanding and influencing human decision-making in social dilemmas.

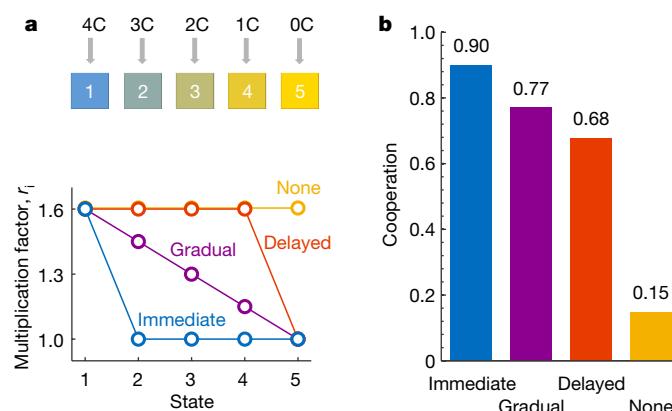


Fig. 4 | Strong immediate feedback maximizes cooperation. **a**, A four-player scenario in which cooperation improves and defection reduces the value of the public good. Transitions are state-independent: the next state depends on only the number of co-operators, not the previous state. In game 1, contributions to a public good are multiplied by the highest factor $r_1 = 1.6$. In game 5, cooperation does not produce any social benefit, $r_5 = c = 1$. For the payoff in the intermediate games 2, 3 and 4, we distinguish three cases: partial defection has immediate, gradual or delayed consequences on the multiplication factor of the public good. In addition, we consider a fourth scenario in which the multiplication factor remains high in all states ('none', no payoff consequences). **b**, An evolutionary analysis confirms that immediately deteriorating public resources are most favourable to cooperation because they make unilateral exploitation a risky strategy. However, all three stochastic games in which the benefits of cooperation vary lead to substantially more cooperation than the game with no environmental feedback.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0277-x>.

Received: 1 November 2017; Accepted: 17 May 2018;
Published online 4 July 2018.

1. Lloyd, W. F. *Two Lectures on the Checks to Population* (Oxford Univ. Press, Oxford, 1833).
2. Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).
3. Trivers, R. L. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
4. Axelrod, R. *The Evolution of Cooperation* (Basic Books, New York, NY, 1984).
5. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge Univ. Press, Cambridge, 1990).
6. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
7. Van Lange, P. A. M., Balliet, D., Parks, C. D. & Van Vugt, M. *Social Dilemmas – The Psychology of Human Cooperation* (Oxford Univ. Press, Oxford, 2015).

8. Sigmund, K. *The Calculus of Selfishness* (Princeton Univ. Press, Princeton, 2010).
9. Nowak, M. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58 (1993).
10. Hauert, C. & Schuster, H. G. Effects of increasing the number of players and memory size in the iterated prisoner's dilemma: a numerical approach. *Proc. R. Soc. Lond. B* **264**, 513–519 (1997).
11. Killingback, T. & Doebeli, M. The continuous prisoner's dilemma and the evolution of cooperation through reciprocal altruism with variable investment. *Am. Nat.* **160**, 421–438 (2002).
12. Szolnoki, A., Perc, M. & Szabó, G. Phase diagrams for three-strategy evolutionary prisoner's dilemma games on regular graphs. *Phys. Rev. E* **80**, 056104 (2009).
13. Grujić, J., Cuesta, J. A. & Sánchez, A. On the coexistence of cooperators, defectors and conditional cooperators in the multiplayer iterated Prisoner's Dilemma. *J. Theor. Biol.* **300**, 299–308 (2012).
14. García, J. & van Veelen, M. In and out of equilibrium I: evolution of strategies in repeated games with discounting. *J. Econ. Theory* **161**, 161–189 (2016).
15. Hilbe, C., Chatterjee, K. & Nowak, M. A. Partners and rivals in direct reciprocity. *Nat. Hum. Behav.* (2018).
16. Shapley, L. S. Stochastic games. *Proc. Natl Acad. Sci. USA* **39**, 1095–1100 (1953).
17. Neyman, A. & Sorin, S. (eds) *Stochastic Games and Applications* (Kluwer Academic Press, Dordrecht, 2003).
18. Mertens, J. F. & Neyman, A. Stochastic games. *Int. J. Game Theory* **10**, 53–66 (1981).
19. Mertens, J. F. & Neyman, A. Stochastic games have a value. *Proc. Natl Acad. Sci. USA* **79**, 2145–2146 (1982).
20. Rand, D. G. & Nowak, M. A. Human cooperation. *Trends Cogn. Sci.* **17**, 413–425 (2013).
21. Ledyard, J. O. in *The Handbook of Experimental Economics* (eds Kagel, J. H. & Roth, A. E.) 111–194 (Princeton Univ. Press, Princeton, 1995).
22. Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A. & Marotzke, J. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proc. Natl Acad. Sci. USA* **105**, 2291–2294 (2008).
23. Alur, R., Henzinger, T. & Kupferman, O. Alternating-time temporal logic. *J. Assoc. Comput. Mach.* **49**, 672–713 (2002).
24. Miltersen, P. B. & Sørensen, T. B. A near-optimal strategy for a heads-up no-limit texas hold'em poker tournament. In *Proc. 6th International Joint Conference on Autonomous Agents and Multiagent Systems* 191 (ACM, 2007).
25. Ashcroft, P., Altrock, P. M. & Galla, T. Fixation in finite populations evolving in fluctuating environments. *J. R. Soc. Interface* **11**, 20140663 (2014).
26. Gokhale, C. S. & Hauert, C. Eco-evolutionary dynamics of social dilemmas. *Theor. Popul. Biol.* **111**, 28–42 (2016).
27. Hauert, C., Holmes, M. & Doebeli, M. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proc. R. Soc. Lond. B* **273**, 2565–2570 (2006); corrigendum 273, 3131–313 (2006).
28. Weitz, J. S., Eksin, C., Paarporn, K., Brown, S. P. & Ratcliff, W. C. An oscillating tragedy of the commons: replicator dynamics with game-environment feedback. *Proc. Natl Acad. Sci. USA* **113**, E7518–E7525 (2016).
29. Tavoni, A., Schlüter, M. & Levin, S. The survival of the conformist: social pressure and renewable resource management. *J. Theor. Biol.* **299**, 152–161 (2012).
30. Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 011909 (2006).

Acknowledgements This work was supported by the European Research Council Start Grant 279307: Graph Games (to K.C.), Austrian Science Fund (FWF) grant P23499-N23 (to K.C.), FWF NFN grant S11407-N23 Rigorous Systems Engineering/Systematic Methods in Systems Engineering (to K.C.), Office of Naval Research Grant N00014-16-1-2914 (to M.A.N.) and the John Templeton Foundation (M.A.N.). C.H. acknowledges support from the ISTFELLOW programme.

Reviewer information *Nature* thanks A. Neyman and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors conceived the study, performed the analysis, discussed the results and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0277-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0277-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.H. or K.C. or M.A.N.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Here we summarize our general framework and the methods that we used. Further details are provided in Supplementary Information.

Stochastic games. To describe a stochastic game fully, we need to specify five objects: (i) the set of players \mathcal{N} , (ii) the set of possible states S , (iii) the set of actions $A(s_i)$ that are available to each player in a given state s_i , (iv) the transition function Q that describes how the current state of the environment and the players' actions in a given round determine the state in the next round, and (v) a payoff function u that describes how the payoffs of the players in a given round depend on the players' actions and on the present state. The framework of stochastic games does not specify how much time passes between consecutive rounds, nor does it restrict the payoffs that are available in each round. The respective model parameters need to be chosen with respect to the specific application (see Supplementary Information for a detailed description of the framework and how it applies to specific examples). Here we have considered scenarios in which players face a strict social dilemma in each state, but the framework can easily be adapted to more general payoff constellations (Extended Data Fig. 10).

Throughout the main text, we considered simple examples of stochastic games. Players can choose between cooperation and defection, and thus their action set is $\{C, D\}$ for each state. Transitions are symmetric: the transition function Q does not depend on which of the players has cooperated or defected. The payoffs per round are symmetric and in the two-player case given by payoff matrices. The payoff of a player in the stochastic game is defined as the player's discounted payoff per round over infinitely many rounds. Initially, players are in state 1. Here we focus on stochastic games that take place in discrete time, but continuous-time stochastic games have also been considered³¹ (see Supplementary Information for a more detailed discussion).

Memory-one strategies. In general, strategies for stochastic games can be arbitrarily complex. A player's action in a given round may depend on the present state and on the whole previous history. To facilitate an evolutionary analysis, we focus on comparably simple strategies^{32–39}: players take into account only the present state and the outcome of the previous round. For n -player games with m states, such 'memory one' strategies can be written as a $2nm$ -dimensional vector $\mathbf{p} = (p_{a,j}^i)$, with $i \in \{1, 2, \dots, m\}$, $j \in \{0, 1, \dots, n-1\}$ and $a \in \{C, D\}$. Each entry $p_{a,j}^i$ represents the player's probability of cooperating in a given round, given that the present state is s_i and that in the previous round the focal player chose action $a \in \{C, D\}$, while j of the $n-1$ other group members cooperated. In Supplementary Table 1, we present several examples of memory-one strategies for stochastic games.

When all players use memory-one strategies, the dynamics of a stochastic game can be described by a Markov chain with $m2^n$ possible states $(s_1, C, \dots, C), \dots, (s_m, D, \dots, D)$. In this notation, the first entry refers to the state of the public good in a given round and the other n entries refer to the players' actions. Using the theory of Markov chains, we compute the players' expected payoffs (see Supplementary Information).

Evolutionary dynamics. To describe how individuals adopt new strategies over time, we consider a standard imitation process³⁰. There is a population of size N . Each member of the population is equipped with a memory-one strategy that prescribes how the individual plays the stochastic game. In each evolutionary time step, every player interacts with every other player to derive a payoff from the stochastic game. Then, two individuals are drawn randomly from the population, a learner and a role model. The payoffs of those two individuals are π_L and π_R , respectively. The learner adopts the strategy of the role model with probability $\rho = 1/[1 + e^{-\beta(\pi_R - \pi_L)}]$. The parameter $\beta \geq 0$ corresponds to the intensity of

selection. For $\beta = 0$, we have random drift. For $\beta > 0$, imitation events are biased in favour of strategies that yield higher payoffs. In addition to imitation events, we allow for random strategy exploration, which corresponds to mutations: with probability μ an individual adopts a randomly chosen memory-one strategy instead of imitating a co-player. We analyse the ergodic mutation-selection process using computer simulations. We obtain exact numerical results when exploration events are rare.

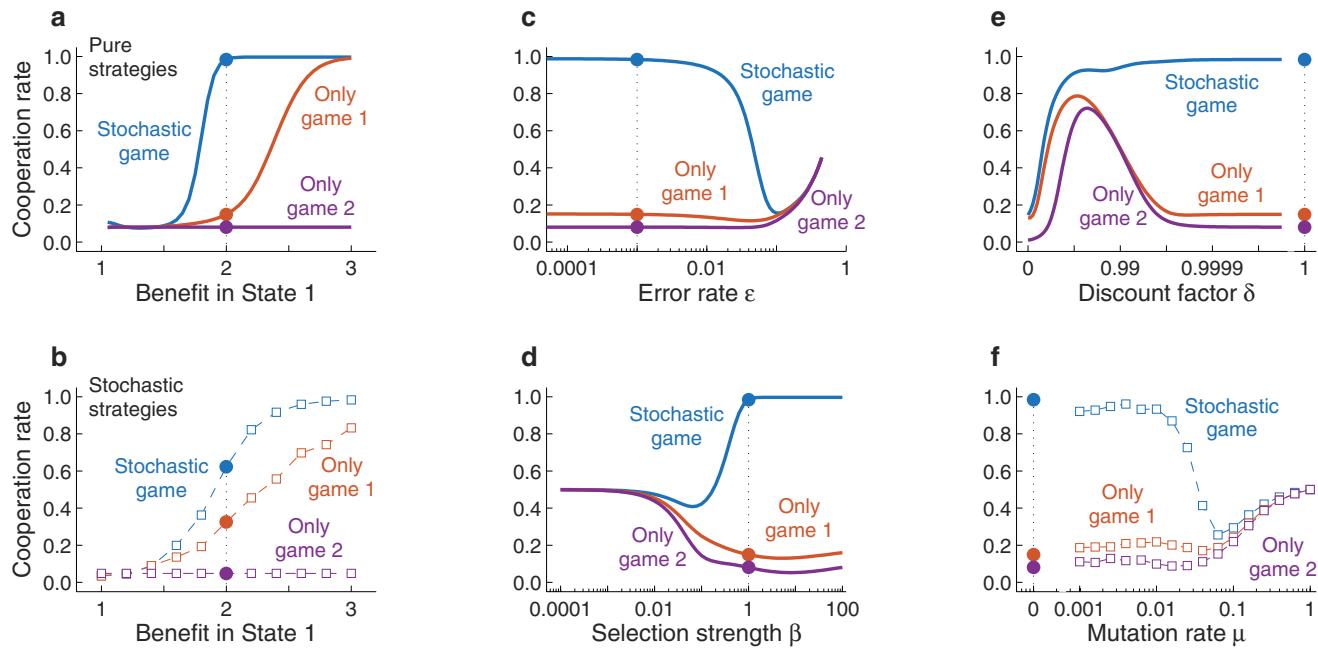
Specific methods used for individual figures. Except for the results in Fig. 3c, the main text considers examples in which players use pure memory-one strategies, subject to small errors (such that $p_{a,j}^i$ is either ε or $1 - \varepsilon$, with $\varepsilon = 0.001$). Further simulations using stochastic memory-one strategies confirm that the respective results are robust (Extended Data Fig. 1b). Except for the stochastic game in Fig. 3b, we assume that future payoffs are not discounted, $\delta \rightarrow 1$. For the evolutionary trajectories of Fig. 2, we averaged over 100 simulations for the scenario with rare mutations. Our numerical results use population size $N = 100$, intermediate selection ($\beta = 1$) for pairwise games and strong selection for multiplayer games ($\beta = 100$ in Fig. 2b and $\beta = 10$ in Fig. 4). Our qualitative findings are robust with respect to parameter changes (Extended Data Fig. 1). For the results in Fig. 3a, b and 4 we report exact results in the limit of rare mutations⁴⁰. Figure 3c shows the phase portrait of adaptive dynamics⁸ for the game with timeout; the corresponding differential equation is derived in Supplementary Information.

Code availability. All simulations and numerical calculations were performed with MATLAB R2014A. In Supplementary Information (see appendix), we provide an algorithm that can be used to calculate payoffs in stochastic games with n players and two states. All other scripts are available from the authors on request or at <https://doi.org/10.5281/zenodo.1287718>.

Data availability. The raw data generated, which were used to create Figs. 2–4, have been uploaded along with the MATLAB code and are available at <https://doi.org/10.5281/zenodo.1287718>.

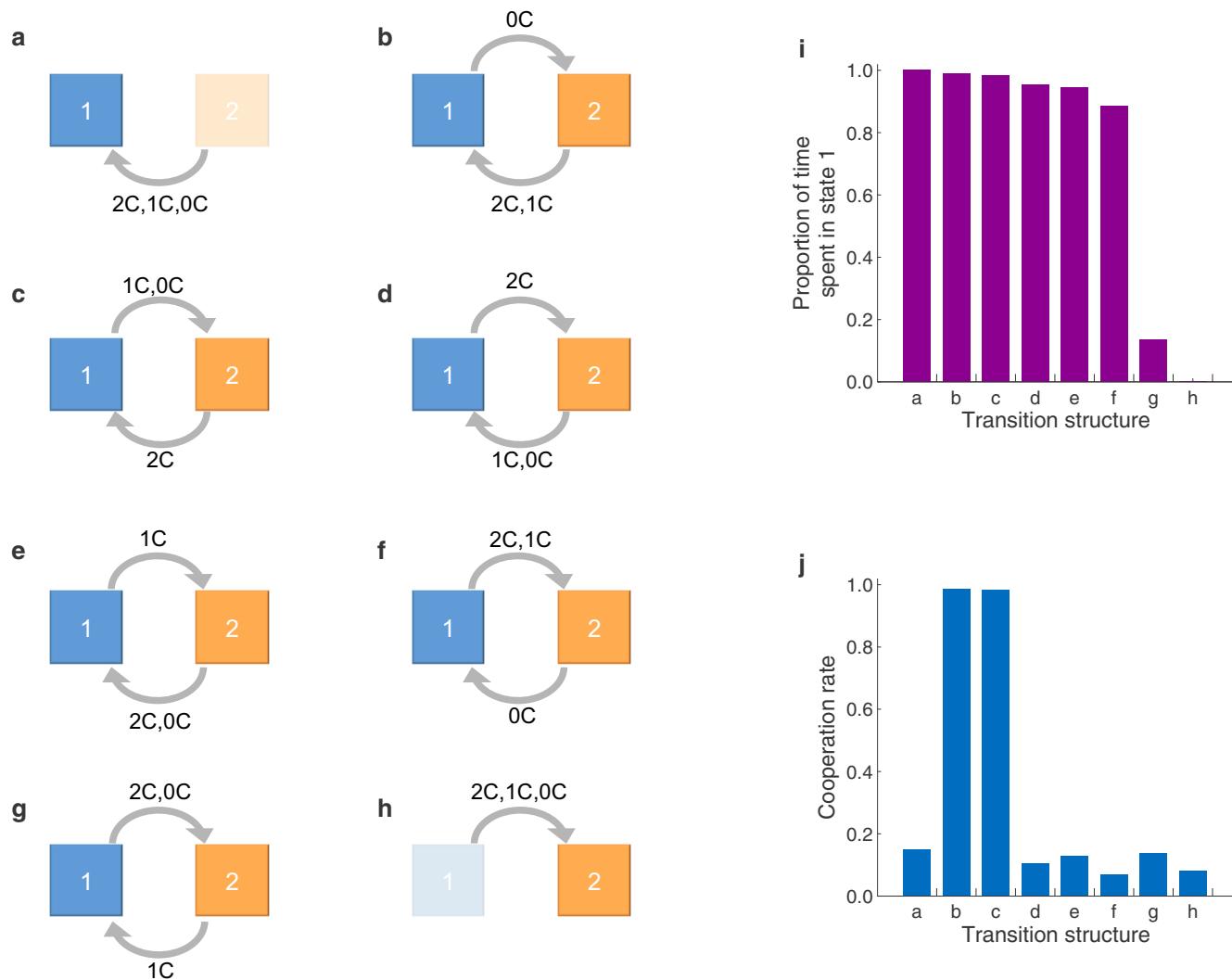
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

31. Neyman, A. Continuous-time stochastic games. *Games Econ. Behav.* **104**, 92–130 (2017).
32. Nowak, M. A. & Sigmund, K. The evolution of stochastic strategies in the prisoner's dilemma. *Acta Appl. Math.* **20**, 247–265 (1990).
33. Ohtsuki, H. & Iwasa, Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
34. Stewart, A. J. & Plotkin, J. B. Collapse of cooperation in evolving games. *Proc. Natl Acad. Sci. USA* **111**, 17558–17563 (2014).
35. Pinheiro, F. L., Vasconcelos, V. V., Santos, F. C. & Pacheco, J. M. Evolution of all-or-none strategies in repeated public goods dilemmas. *PLOS Comput. Biol.* **10**, e1003945 (2014).
36. Akin, E. in *Ergodic Theory, Advances in Dynamics* (ed. Assani, I.) 77–107 (de Gruyter, Berlin, 2016).
37. Hilbe, C., Martinez-Vaquero, L. A., Chatterjee, K. & Nowak, M. A. Memory- n strategies of direct reciprocity. *Proc. Natl Acad. Sci. USA* **114**, 4715–4720 (2017).
38. Stewart, A. J. & Plotkin, J. B. Small groups and long memories promote cooperation. *Sci. Rep.* **6**, 26889 (2016).
39. Reiter, J. G., Hilbe, C., Rand, D. G., Chatterjee, K. & Nowak, M. A. Crosstalk in concurrent repeated games impedes direct reciprocity and requires stronger levels of forgiveness. *Nat. Commun.* **9**, 555 (2018).
40. Fudenberg, D. & Imhof, L. A. Imitation processes with small mutations. *J. Econ. Theory* **131**, 251–262 (2006).



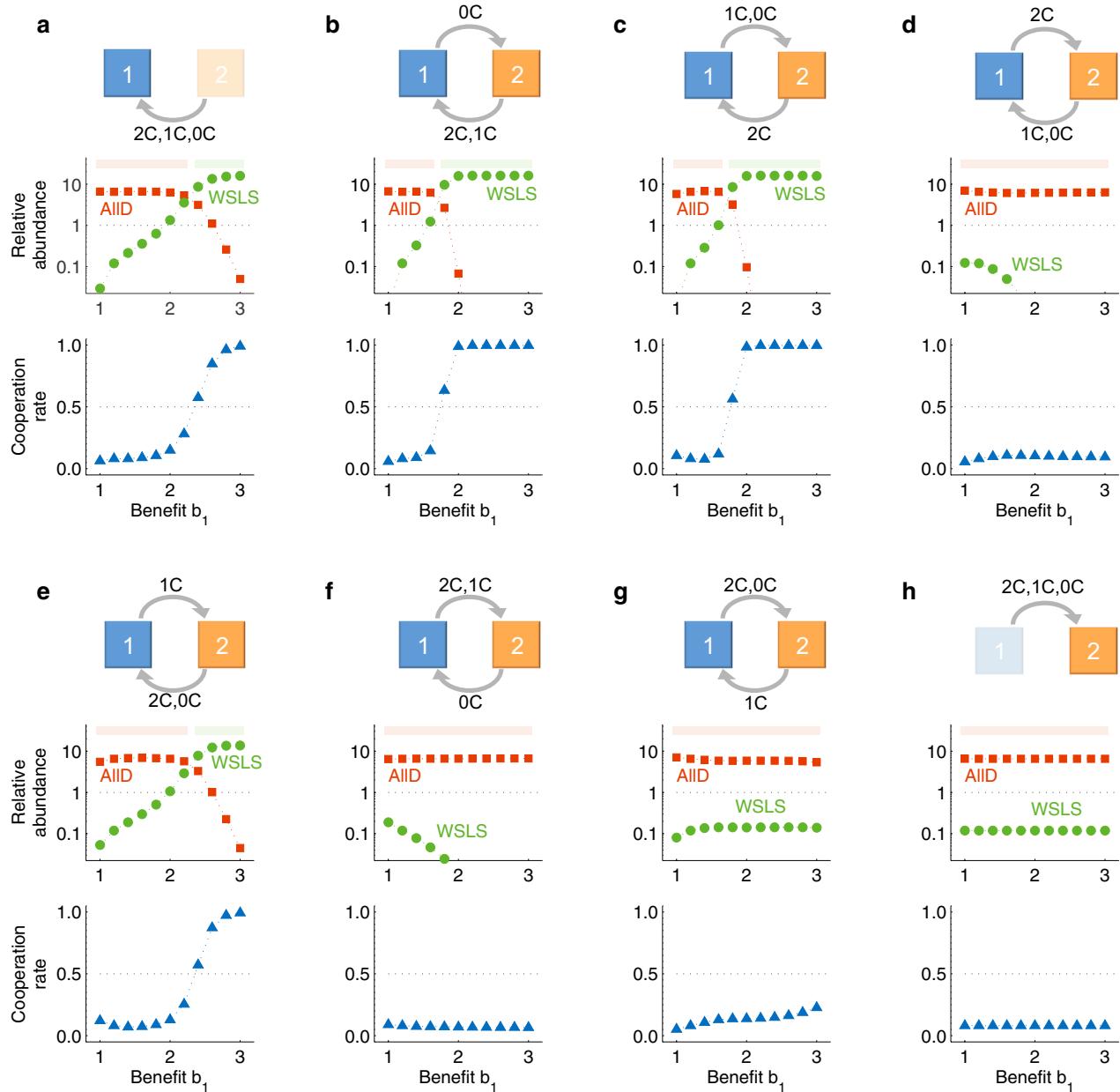
Extended Data Fig. 1 | Our findings are robust with respect to parameter changes. To test the robustness of our findings, we consider the stochastic game introduced in Fig. 2a and independently vary several key parameters. **a, b**, When we vary the benefit of cooperation in state 1, we find that the advantage of the stochastic game is most pronounced when this benefit is intermediate, $1.5 \leq b_1 \leq 2.5$. This conclusion holds independently of whether individuals use pure strategies only (**a**) or stochastic ones (**b**). **c-f**, We obtain similar results when we vary the error rate ε (**c**), the strength of selection β (**d**), the discount factor δ (**e**) and the

mutation rate μ (**f**). In all cases, we observe that stochastic games yield a cooperation premium, provided that errors are sufficiently rare, selection is sufficiently strong, players give sufficient weight to future payoffs and mutations are comparably rare. Solid lines indicate exact results in the limit of rare mutations, whereas square symbols and dashed lines represent simulation results (see Supplementary Information for details). Filled circles highlight the results obtained for the parameters in Fig. 2a. As default parameters, we used the same values as in Fig. 2a: $N = 100$, $b_1 = 2.0$, $b_2 = 1.2$, $c = 1$, $\beta = 1$, $\varepsilon = 0.001$, $\delta \rightarrow 1$ and $\mu \rightarrow 0$.



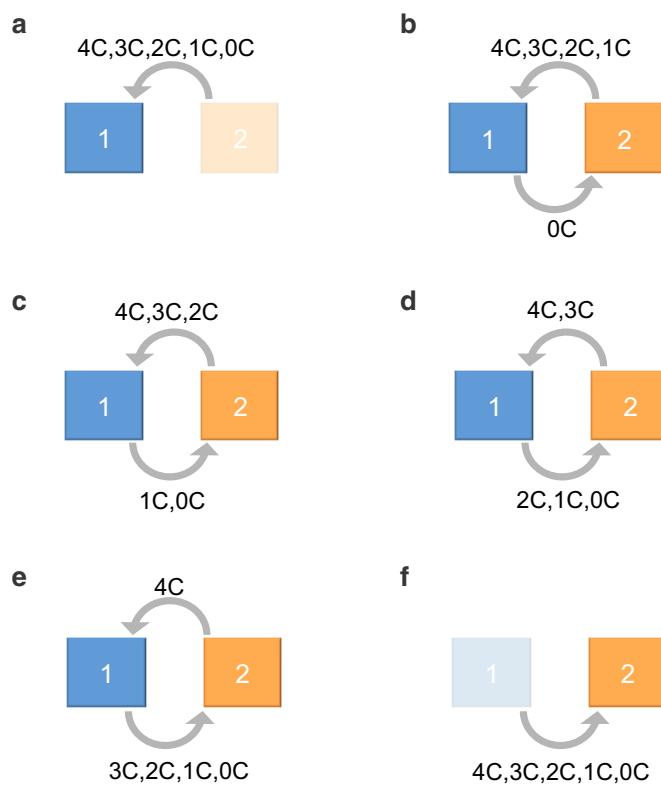
Extended Data Fig. 2 | Whether cooperation evolves in two-player games depends critically on the form of the environmental feedback. Keeping the game parameters fixed at the values used in Fig. 2a, we explored how the evolution of cooperation depends on the underlying transition structure of the stochastic game in the limit of rare mutations (see Supplementary Information). **a–h**, We calculated the selection-mutation equilibrium for all possible stochastic games with two states when transitions are state-independent and deterministic. **i**, Overall, six of the eight transition structures lead players to spend more time in the

more profitable state 1, in which mutual cooperation has a higher benefit. **j**, However, cooperation evolves in only two out of these six transition structures. These two structures have in common that mutual cooperation always leads to the beneficial state 1, whereas mutual defection leads to the detrimental state 2. Thus, cooperation is most likely to evolve if the environmental feedback itself incentivizes mutual cooperation and disincentivizes mutual defection. The transitions after unilateral defection have a less prominent role.

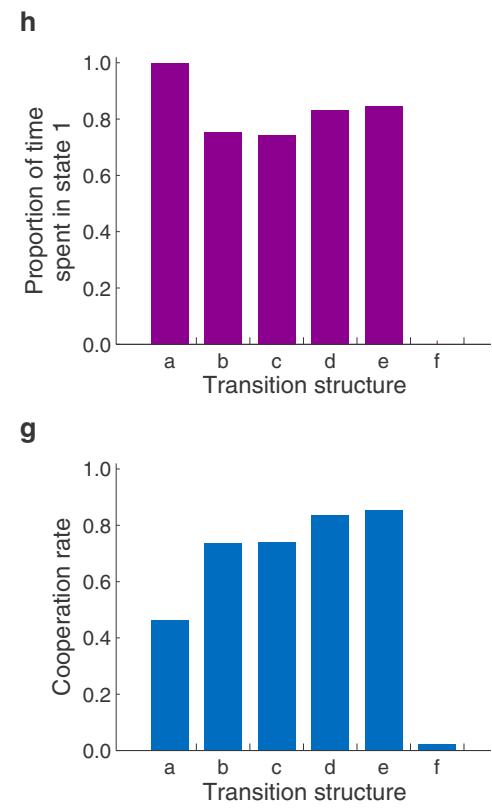


Extended Data Fig. 3 | Analysis of the evolving strategies suggests that the evolution of cooperation hinges on the success of WSLS. Here, we consider all state-invariant and deterministic stochastic games with two states and two players. **a-h**, For each of the eight possible cases, we recorded the evolving cooperation rate (lower plots) and the relative abundance of each pure memory-one strategy (upper plots) for different values of b_1 . For clarity, we depict only two memory-one strategies explicitly, All D (the strategy that prescribes to always defect) and WSLS. The colour-shaded bars on top of the upper plots show parameter regimes

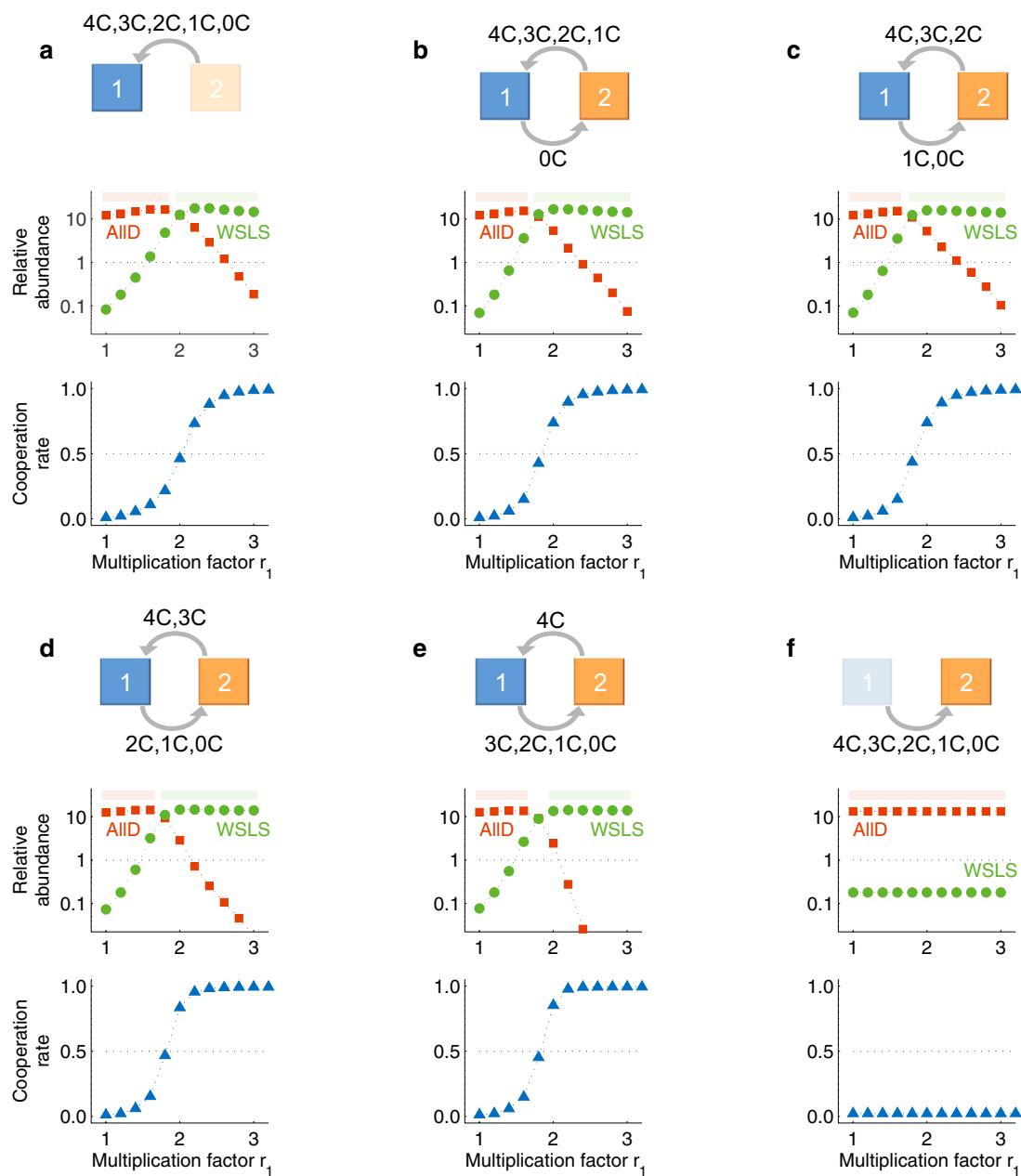
in which either All D or WSLS is most abundant among all 16 strategies. In four of the eight cases, we observe that full cooperation evolves as the benefit to cooperation in state 1 approaches $b_1 = 3$. These are exactly the cases in which mutual cooperation leads players towards the more beneficial state 1. Moreover, in these four cases the upper plots show that cooperation emerges owing to the success of WSLS, which is the predominant strategy whenever cooperation prevails. Except for the value of b_1 , all other parameter values are the same as in Extended Data Fig. 2.



Extended Data Fig. 4 | Effect of transitions on cooperation in four-player public-goods games. We also explored the effect of different transition structures for stochastic games between multiple players (with a public-goods game being played in each state). State 1 is again more beneficial because $r_1 > r_2$, but to be in state 1 there must be a minimum number k of cooperators in the previous round. **a–f**, For a four-player public-goods game, there are six possible monotonic configurations of the stochastic game because k can be any number from 0 (players always

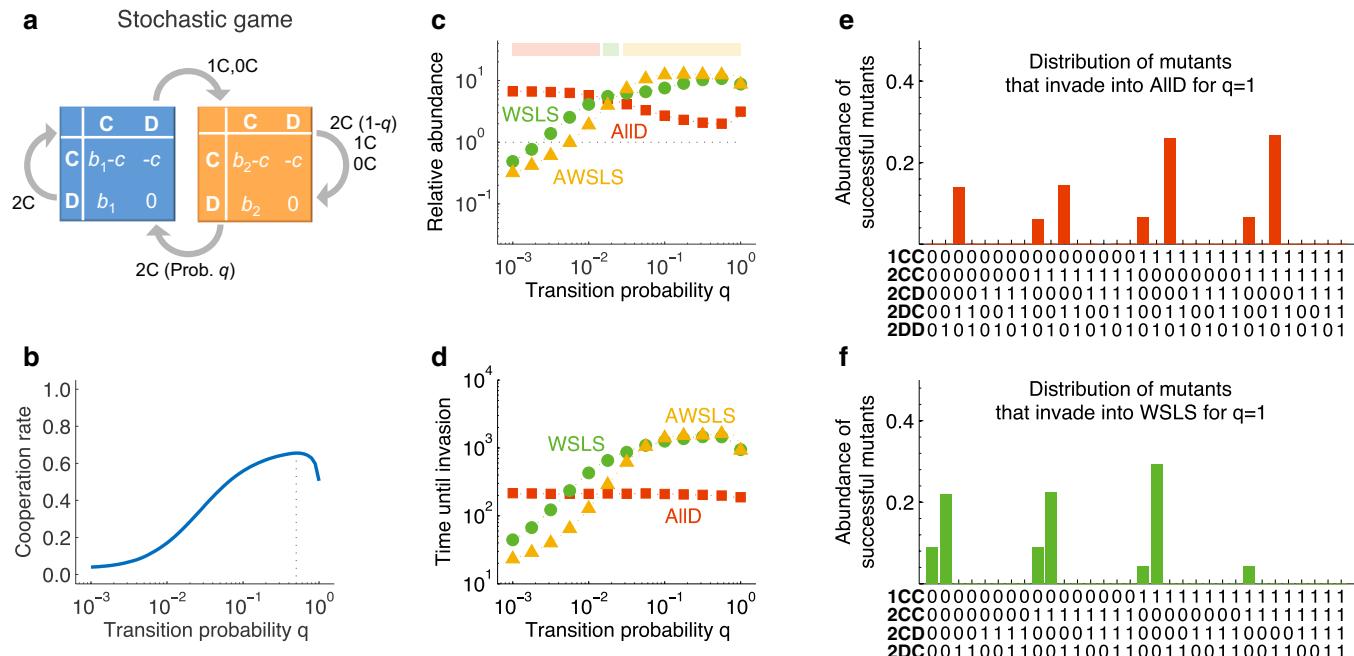


move to first state) to 5 (players never move to first state). **h**, There is a non-monotonic relationship between the six transition structures and the time spent in the more beneficial state 1. **g**, The evolving cooperation rate becomes maximal when any deviation from mutual cooperation leads players to state 2 (e). Parameters are as in Fig. 2b, but with the multiplication factor in the first state fixed to $r_1 = 2$ and selection strength $\beta = 1$; to derive exact results, we considered the limit of rare mutations $\mu \rightarrow 0$ (see Supplementary Information for details).



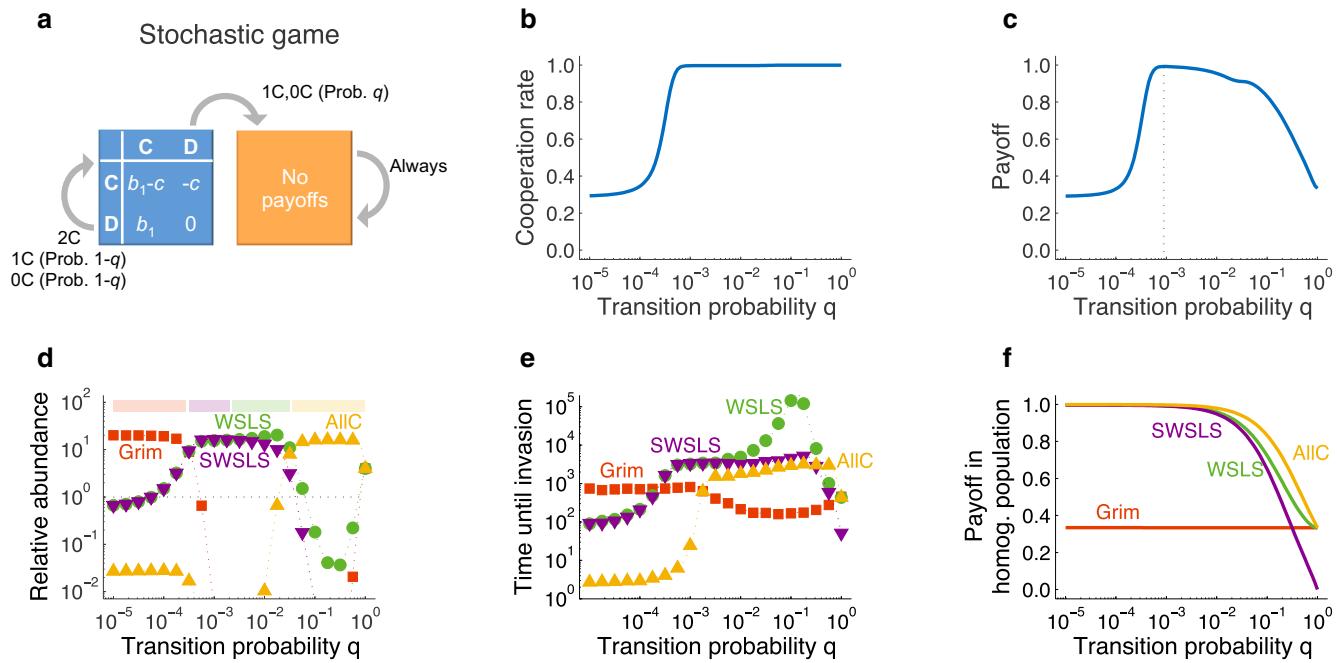
Extended Data Fig. 5 | WSLS sustains cooperation in multiplayer public-goods games. This figure is analogous to Extended Data Fig. 3 for the case of multiplayer interactions. Again, we show evolving cooperation rates and the relative abundance of All D and WSLS for the six state-independent and deterministic games in which transitions are monotonic.

In five of these games, cooperation emerges once the multiplication factor r_1 becomes sufficiently large. In all of those, WSLS is the most abundant strategy when cooperation evolves. Except for r_1 , all parameters are the same as in Extended Data Fig. 4.



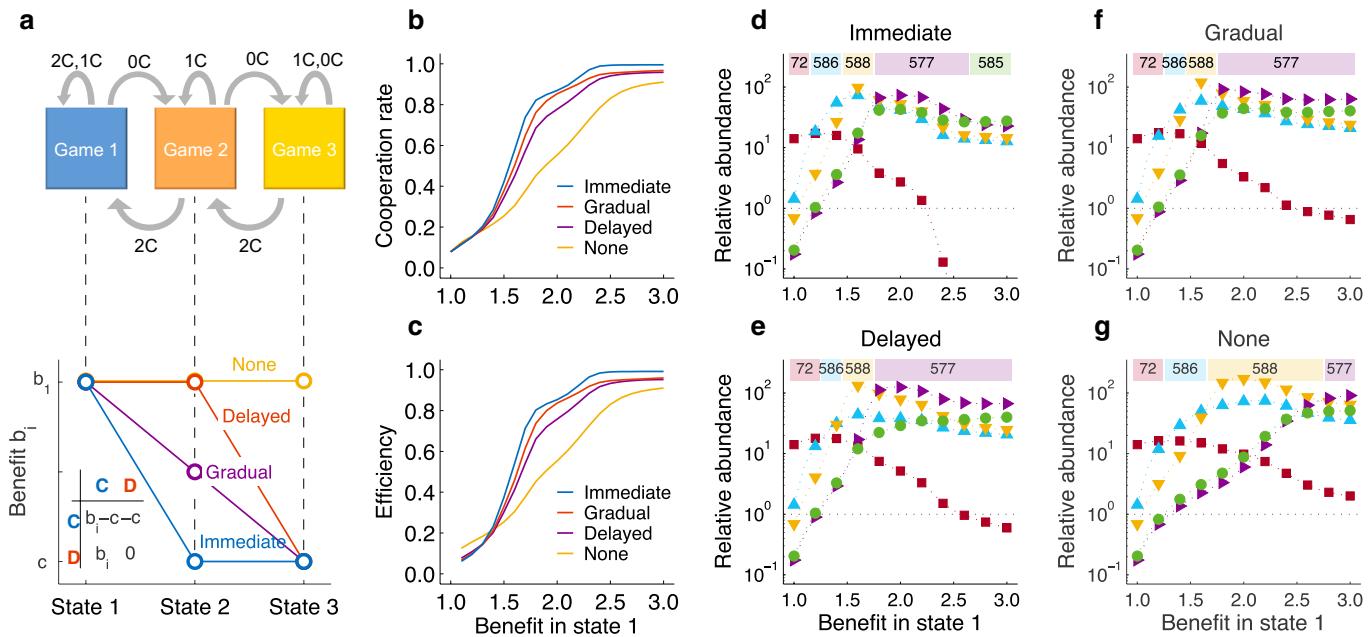
Extended Data Fig. 6 | Probabilistic transitions can further enhance cooperation. **a**, Here, we explore in more detail the stochastic game introduced in Fig. 3a (see Supplementary Information for details), in which any defection always leads to state 2. After mutual cooperation in state 1, players remain in state 1 with certainty. After mutual cooperation in state 2, players move towards state 1 with probability q . **b**, Calculating the cooperation rate in the selection–mutation equilibrium in the limit of rare mutations shows that the highest cooperation rate is achieved for intermediate values of q . **c**, We recorded the abundance of all 32 memory-one strategies in the selection–mutation equilibrium. The most abundant strategy is either All D (for small values of q , as indicated by

the red squares), WSLS (for small but positive values of q , green circles) or AWSLS (for all other values of q , yellow triangles; AWSLS is a more ambitious variant of WSLS, see Supplementary Information, section 4.1). **d**, To estimate the time that it takes each resident strategy to be invaded, we randomly introduced other mutant strategies and recorded how long it took until a mutant successfully fixed (that is, the number of independent mutant strategies introduced before the mutant strategy was adopted by the whole population). To obtain a reliable estimate, we performed 10,000 runs for each resident strategy. **e**, **f**, In addition, we recorded which strategy eventually reaches fixation if the resident applies either All D or WSLS when $q = 1$. Parameters: $b_1 = 1.9$, $b_2 = 1.4$, $c = 1$, $\beta = 1$, $N = 100$.



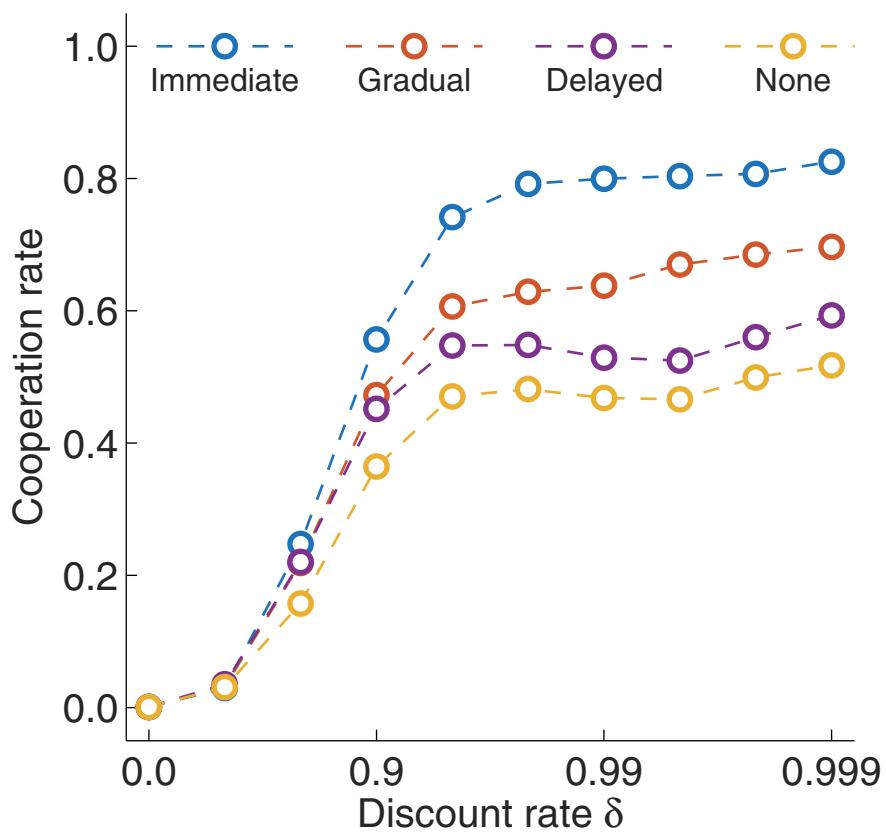
Extended Data Fig. 7 | Players benefit from a small endogenous risk that the game stops early. **a**, We consider the stochastic game in Fig. 3b, in which players remain in state 1 after cooperation, but move towards state 2 with transition probability q if one of the players defects. In state 2, no profitable interactions are possible. All results are discussed in detail in Supplementary Information; here we provide a summary. **b**, According to our evolutionary simulations, a higher transition probability leads to more cooperation. **c**, However, a higher probability q also makes players move to the second state if one of them defected merely owing to an error; hence, the dependence of payoffs on q is non-monotonic. **d, e**, When

q is small, Grim is the predominant strategy. Players with this strategy cooperate until one of the players defects; from then on, they defect forever. As q increases, WSLS strategies take over. As $q \rightarrow 1$, unconditional cooperation becomes most successful. **f**, For the given parameter values, a homogeneous Grim population achieves only one-third of the maximum payoff possible, because any error leads to relentless defection. The other three strategies result in the maximum payoff $b_1 - c$ for $q = 0$, but this payoff decreases with q . Parameters: $b_1 = 2$, $c = 1$, $\delta = 0.999$, $\varepsilon = 0.001$, $\beta = 1$, $N = 100$.



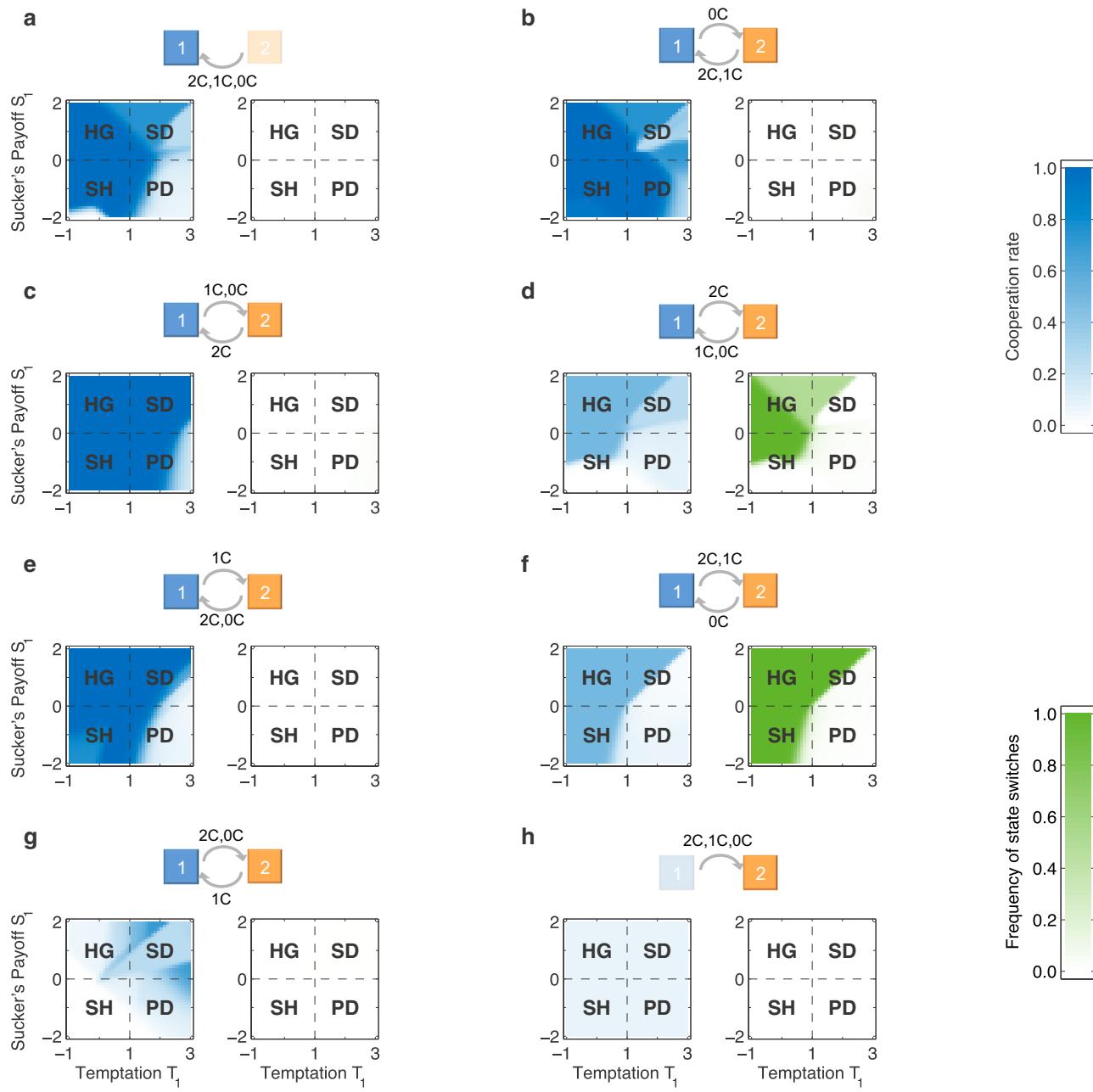
Extended Data Fig. 8 | Immediate environmental feedback enhances cooperation. **a**, We consider a state-dependent stochastic game with two players and three states. Mutual cooperation always leads players to move to a superior state (or to remain in the most beneficial state s_1). Similarly, mutual defection always leads to an inferior state (or players remain in the most detrimental state s_3). After a unilateral defection, players remain in the same state. We consider four different versions of this game, depending on how quickly the payoffs decrease as players move towards an inferior state. **b**, Our numerical results show that an immediate negative response of the environment to defection is most favourable to the evolution of cooperation. **c**, As a consequence, the scenario with immediate

consequences also yields the highest average payoffs once the benefit in state 1 exceeds a moderate threshold. **d–g**, On the level of evolving strategies, we find that an immediately responding environment is most favourable to the evolution of WSLS strategies and strongly selects against defecting strategies. Again, the coloured bars on top of each panel indicate the strategy that is most favoured by selection for the respective value of b_1 (see Supplementary Information for all details). Parameters: $c = 1$; b_1 varies from 1 to 3; b_2 is equal to c , $(b_1 + c)/2$ or b_1 ; and b_3 is equal to either c or b_1 depending on the scenario considered (as depicted in **a**); $N = 100$, $\beta = 1$, $\delta \rightarrow 1$, $\varepsilon = 0.001$.



Extended Data Fig. 9 | Cooperation in stochastic games requires that players take future payoff consequences into account. We repeated the numerical computations in Extended Data Fig. 8 for various discount rates δ . When players focus entirely on the present ($\delta = 0$), cooperation evolves in none of the four treatments. As players increasingly take future payoffs

into account, cooperation rates increase. Immediate payoff feedback is most conducive to cooperation across all values of δ considered. Except for the discount rate, parameters are the same as in Extended Data Fig. 8, with $b_1 = 1.8$.



Extended Data Fig. 10 | A systematic analysis of the expected game dynamics for different game payoffs. Keeping the two-player game in state 2 fixed to the game in Fig. 2a, we varied the game that is played in state 1. We assume that payoffs in the first state are 1 (for mutual cooperation), S_1 (for unilateral cooperation), T_1 (for unilateral defection) and 0 (for mutual defection). Depending on T_1 and S_1 , game 1 can be one of four different types: harmony game (HG), snowdrift game (SD), stag-hunt game (SH) or prisoner's dilemma (PD); see Supplementary Information for details. For each of the eight possible state-independent transitions q , we systematically varied the temptation payoff T_1 (x axis) and the sucker's payoff S_1 (y axis) in the first state (see Supplementary

Information for details). For each combination of T_1 , S_1 and q , we computed how often players cooperate in the selection-mutation equilibrium (left panels) and in what fraction of rounds they switch from one state to the other (right panels). **a-c, e**, Full cooperation can evolve when players find themselves in state 1 after mutual cooperation. **d, f**, Players learn to switch between states only when mutual cooperation leads to state 2 and mutual defection leads to state 1. **g, h**, In the remaining cases, players hardly cooperate. The payoffs in game 2 are the same as in Fig. 2a—a prisoner's dilemma with $b_2 = 1.2$ and $c = 1$. For the evolutionary parameters we considered population size $N = 100$ and selection strength $\beta = 1$.

Seabirds enhance coral reef productivity and functioning in the absence of invasive rats

Nicholas A. J. Graham^{1,2*}, Shaun K. Wilson^{3,4}, Peter Carr^{5,6}, Andrew S. Hoey², Simon Jennings⁷ & M. Aaron MacNeil⁸

Biotic connectivity between ecosystems can provide major transport of organic matter and nutrients, influencing ecosystem structure and productivity¹, yet the implications are poorly understood owing to human disruptions of natural flows². When abundant, seabirds feeding in the open ocean transport large quantities of nutrients onto islands, enhancing the productivity of island fauna and flora^{3,4}. Whether leaching of these nutrients back into the sea influences the productivity, structure and functioning of adjacent coral reef ecosystems is not known. Here we address this question using a rare natural experiment in the Chagos Archipelago, in which some islands are rat-infested and others are rat-free. We found that seabird densities and nitrogen deposition rates are 760 and 251 times higher, respectively, on islands where humans have not introduced rats. Consequently, rat-free islands had substantially higher nitrogen stable isotope ($\delta^{15}\text{N}$) values in soils and shrubs, reflecting pelagic nutrient sources. These higher values of $\delta^{15}\text{N}$ were also apparent in macroalgae, filter-feeding sponges, turf algae and fish on adjacent coral reefs. Herbivorous damselfish on reefs adjacent to the rat-free islands grew faster, and fish communities had higher biomass across trophic feeding groups, with 48% greater overall biomass. Rates of two critical ecosystem functions, grazing and bioerosion, were 3.2 and 3.8 times higher, respectively, adjacent to rat-free islands. Collectively, these results reveal how rat introductions disrupt nutrient flows among pelagic, island and coral reef ecosystems. Thus, rat eradication on oceanic islands should be a high conservation priority as it is likely to benefit terrestrial ecosystems and enhance coral reef productivity and functioning by restoring seabird-derived nutrient subsidies from large areas of ocean.

The flow of organic matter and nutrients among ecosystems is a major determinant of productivity, composition and functioning. Animals, such as moose⁵, salmon⁶ and sea turtles⁷, can connect ecosystems by vectoring organic matter and nutrients between them. However, the magnitude and implications of these natural dynamics are poorly understood in contemporary ecosystems in which humans have disrupted connectivity by creating barriers such as dams, removing biomass and introducing predators^{2,8}. Seabirds are globally important drivers of nutrient cycling⁹, transferring nutrients from their pelagic feeding grounds to islands on which they roost and breed^{1,10}. This input of nutrient-rich guano increases plant biomass, alters species compositions of island plants, and enhances the abundance of many types of biota^{3,4}. Nutrients can leach from guano to adjacent marine systems, which may bolster plankton densities and influence feeding behaviour of manta rays^{11,12}. However, the effects of seabird-transported nutrients on the productivity, structure, and function of highly diverse coral reefs are currently unknown. Understanding natural nutrient connectivity is particularly important, yet challenging, because invasive predators such as rats and foxes have decimated seabird populations within 90% of the world's temperate and tropical island groups⁸.

Here we isolate the effects of seabird-derived nutrients on adjacent coral reefs using a rare, large-scale natural experiment in which some

islands in a remote coral reef archipelago are rat-infested, whereas others are rat-free. The northern atolls of the Chagos Archipelago, located in the central Indian Ocean, have been uninhabited by people for over 40 years, are protected from fishing, and host some of the world's most pristine marine environments¹³. Black rats (*Rattus rattus*) are thought to have been introduced to the archipelago in the late 18th and early 19th centuries, but owing to patterns of human habitation and movement, are not present on all islands. We use this unique scenario and a mixed-methods approach to investigate nutrient flux between oceanic, island, and coral reef ecosystems.

We studied six rat-free and six rat-infested islands, selected to be otherwise similar in terms of size, location and environment. Rats are known to predate upon bird eggs, chicks, and occasionally adults, decimating populations where they have been introduced⁸. Mean seabird density, averaged across a six-year period (Methods 'Seabird surveys'), on rat-free islands was 760 times greater than on rat-infested islands (Fig. 1a; 1,243 birds per ha rat-free, 1.6 birds per ha rat-infested). Owing to the high seabird densities on some islands, the Chagos Archipelago has ten Important Bird and Biodiversity Areas¹⁴. The biomass of 14 bird species within six families varied among islands, with terns and noddies contributing the most biomass, and boobies, shearwaters and frigate birds only common on some islands. Biomass of all species was greatest on rat-free islands (Fig. 1b).

We used species-specific abundance, body size-scaled defecation rate, nitrogen content of guano¹⁵, and mean residence times on the islands to estimate mean nitrogen input by the seabirds (Methods 'Seabird surveys'). The nitrogen input by seabirds per hectare of island was 251 times greater on rat-free islands than on rat-infested islands (Fig. 1c; 190 kg ha⁻¹ yr⁻¹ rat-free, 0.8 kg ha⁻¹ yr⁻¹ rat-infested). The nutrient input onto rat-free islands is comparable to nitrogen inputs by seabirds at the isolated Palmyra atoll in the Pacific Ocean¹⁵. We did not calculate nutrient input from rats as they are recycling nutrients already present on the islands. By contrast, the majority of the seabirds feed in the open ocean, substantial distances from reefs (Extended Data Table 1). By foraging offshore, seabirds feed from food webs supported by net primary production that is estimated to be 2–5 orders of magnitude higher than net primary production on adjacent coral reefs (Methods and Extended Data Fig. 1). Their capacity to access these oceanic prey resources leads to substantial deposition of oceanic nitrogen that would otherwise be unavailable on rat-free islands.

We used the abundance of nitrogen and stable isotopes (reported as δ values for the ratio of $^{15}\text{N} : ^{14}\text{N}$ ($\delta^{15}\text{N}$)) to understand the uptake of nutrients on islands and in adjacent coral reef ecosystems (Methods 'Isotope sampling' and Fig. 2). Total nitrogen and $\delta^{15}\text{N}$ were strongly and positively correlated ($r = 0.96$), meaning that they show similar patterns in our samples. Soils on rat-free islands were enriched in ^{15}N , with $\delta^{15}\text{N}$ being 3.8 times higher than on rat-infested islands and comparable to reported values for seabird guano¹⁶ (Fig. 2b). Substantially greater $\delta^{15}\text{N}$ was also evident in new growth leaves of a coastal plant

¹Lancaster Environment Centre, Lancaster University, Lancaster, UK. ²ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland, Australia. ³Department of Biodiversity, Conservation and Attractions, Perth, Western Australia, Australia. ⁴Oceans Institute, University of Western Australia, Crawley, Western Australia, Australia. ⁵Institute of Zoology, Zoological Society of London, London, UK. ⁶College of Life and Environmental Sciences, University of Exeter, Exeter, UK. ⁷International Council for the Exploration of the Sea, Copenhagen, Denmark. ⁸Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada. *e-mail: nick.graham@lancaster.ac.uk

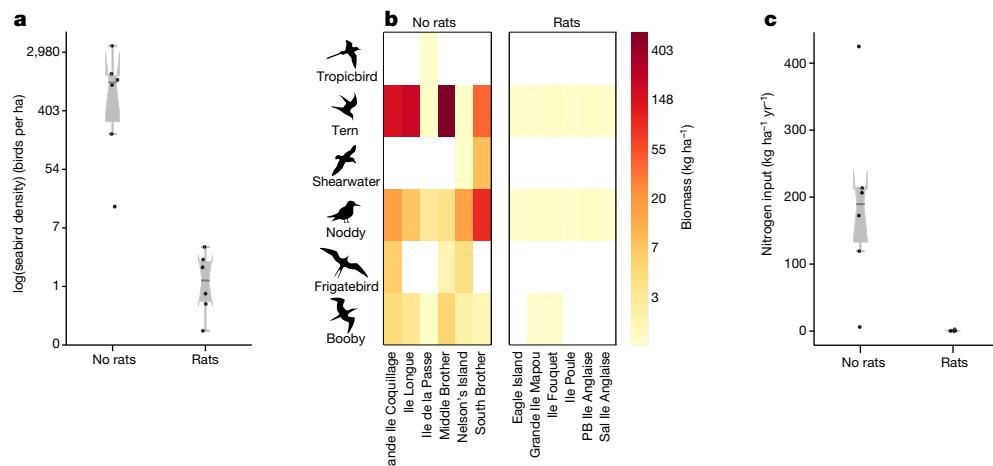


Fig. 1 | Seabird densities, biomass and nitrogen input to islands with and without rats in the Chagos Archipelago. **a**, Seabird density on rat-free ($n=6$) and rat-infested ($n=6$) islands. **b**, Heat maps of seabird biomass per family, on each island. Tropicbird: *Phaethon lepturus*; tern: *Thalasseus bergii*, *Sterna sumatrana*, *Sterna dougallii*, *Onychoprion fuscatus*, *Onychoprion anaethetus*, *Gygis alba*; shearwater: *Puffinus bailloni*, *nicolae*, *Ardenna pacifica*; noddy: *Anous tenuirostris*, *Anous stolidus*; frigatebird: *Fregata* spp.; booby: *Sula sula*, *Sula leucogaster*.

(*Scaevola taccada*) on rat-free islands (Fig. 2c), indicating uptake of oceanic-derived nutrients by island vegetation.

Nitrogen is expected to leach off islands to nearshore marine environments through rainfall and coastal advection¹¹. On the reef flat (approximately 1 m deep and 100 m from the shore) filter-feeding sponges (*Spheciopspongia* sp.; Fig. 1d) and macroalgae (*Halimeda* sp.; Fig. 1e) had substantially higher $\delta^{15}\text{N}$ values near rat-free islands, although differences were smaller than observed for island soils and vegetation. This is consistent with findings of higher $\delta^{15}\text{N}$ values in corals closer to seabird colonies in New Caledonia¹⁷. On the reef crest (approximately 3 m deep and 230 ± 55 m (mean \pm s.d.) from island shorelines) $\delta^{15}\text{N}$ was substantially higher in turf algae and the muscle of herbivorous damselfish (*Plectroglyphidodon lacrymatus*) adjacent to rat-free islands (Fig. 2f, g). While recognizing the influence of trophic fractionation on $\delta^{15}\text{N}$ signatures, the relative depletion of the heavy isotope ^{15}N from the soils on rat-free islands across to the reef crests,

PB, Peros Banhos atoll; Sal, Salomon atoll. **c**, Nitrogen input by seabirds per hectare for rat-free ($n=6$) and rat-infested ($n=6$) islands. **a**, **c**, Notched box plots, in which the horizontal line is the median, box height depicts the interquartile range, whiskers represent 95% quantiles, and diagonal notches illustrate approximate 95% confidence intervals around the median. Estimated net rat effects (median and 95% highest posterior density intervals) are: **a**, 456 [22, 6393] birds per ha; **b**, 195 [184, 207] kg ha $^{-1}$ (total biomass) and **c**, 148 [81, 211] kg ha $^{-1}$ yr $^{-1}$.

compared to the relatively stable values for rat-infested islands, provides strong evidence of seabird-vectored nutrient enrichment propagating out onto adjacent coral reefs. The diminishing effect sizes from the islands out to the reef crest probably reflect a range of processes, including uptake and conversion of nitrogen by micro- and macroorganisms across the reef flat¹⁸.

Comparison of damselfish growth on reef crests (using growth bands in otoliths; Methods 'Fish growth') demonstrated that individuals adjacent to rat-free islands were growing significantly faster towards their maximum expected size ($K_r - K = -0.10 [-0.18, -0.04]$ (95% highest posterior density intervals), net rat effect), and were larger for a given age than individuals on reefs adjacent to rat-infested islands (Fig. 3). This is the first evidence, to our knowledge, for seabird-vectored nutrient subsidies propagating through the food web to accelerate the growth of a marine vertebrate. Given the diversity and high biomass of fishes that feed on benthic algae on coral reefs¹⁹, this finding is likely to

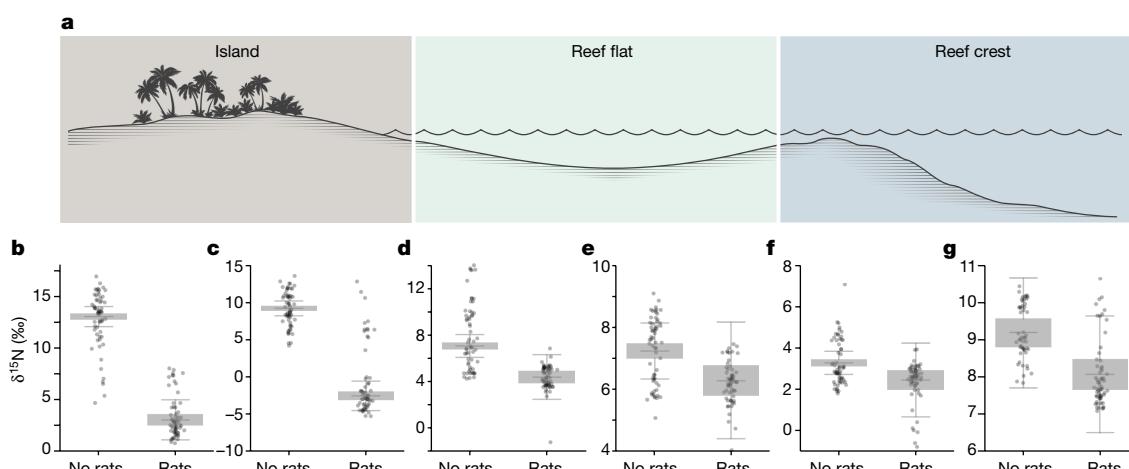


Fig. 2 | Nitrogen isotope signals from islands to reefs in the presence and absence of invasive rats. **a**, Schematic of study system. **b–g**, $\delta^{15}\text{N}$ values for soil (b) and new growth leaves (*S. taccada*) on islands (c), filter feeding sponges (*Spheciopspongia* sp.) (d) and macroalgae (*Halimeda* sp.) on reef flats (e), and turf algae (f) and dorsal muscle tissue of damselfish (*P. lacrymatus*) on reef crests (g). For all groups 120 samples were collected, except for g, for which 110 samples were collected (Methods 'Isotope sampling').

For box plots, the horizontal line is the median, box height depicts first and third quartiles and whiskers represent the 95th percentile. Net rat effect (median [95% highest posterior density]) and the probability of the effect being less than zero ($P(\text{neg})$) estimates are: **b**, $-9.9 [-11.3, -8.4]$, $P(\text{neg}) > 0.99$; **c**, $-11.8 [-13.2, -10.2]$, $P(\text{neg}) > 0.99$; **d**, $-1.0 [-2.3, 0.5]$, $P(\text{neg}) = 0.92$; **e**, $-2.7 [-4.1, -1.23]$, $P(\text{neg}) > 0.99$; **f**, $-0.8 [-2.23, 0.6]$, $P(\text{neg}) = 0.90$; **g**, $-1.1 [-2.5, 0.3]$, $P(\text{neg}) = 0.94$.

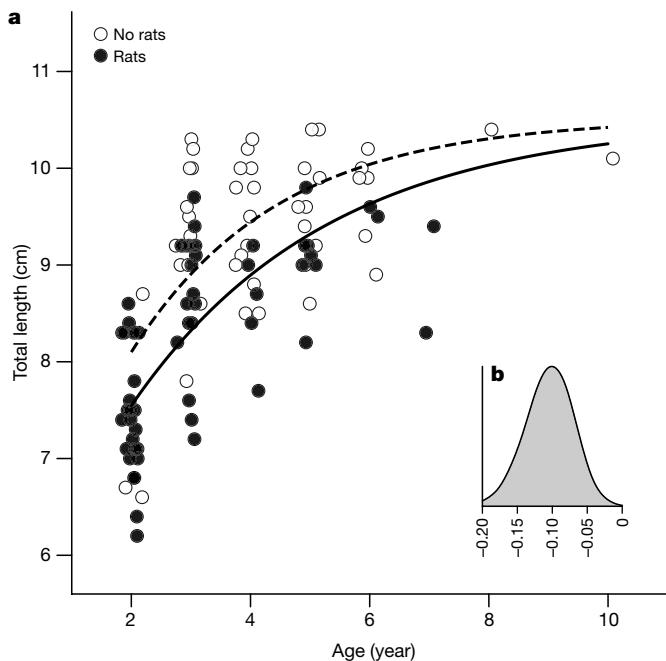


Fig. 3 | Growth of herbivorous damselfish on coral reefs adjacent to islands with and without rats. **a**, Age-by-length growth curves for *P. lacrymatus* on rat-free (open circles) and rat-infested (closed circles) islands. **b**, Effect-size posterior density for the difference between the growth parameter, K (Yr^{-1}), on rat-free compared to rat-infested islands. $n = 48$ and $n = 58$ biologically independent samples for rat-free and rat-infested islands, respectively.

indicate higher fish production adjacent to seabird-dominated islands with repercussions for the production of their predators.

To assess the influence of seabird colonies on reef-fish biomass production, we surveyed fish communities along the reef crests of the islands (Methods 'Fish biomass and function'). Total biomass of the reef-fish community was 48% greater adjacent to rat-free islands.

Assigning the 123 species of reef fish recorded into feeding groups, we found biomass to be greater for all feeding groups of fish on reefs adjacent to rat-free islands, with herbivore biomass having the largest effect size (93% of posterior distribution above zero; Fig. 4a). These results are consistent with seabird-vectored nutrients subsidising the entire ecosystem.

Herbivorous fish are functionally important on coral reefs, maintaining a healthy balance between corals and algae, and clearing space for coral settlement²⁰. Parrotfishes are among the most abundant and important herbivorous groups, providing unique grazing and bio-erosion functions. We estimated grazing and bioerosion rates of parrotfishes for each island using density data, along with species- and body size-specific information on consumption rates²¹ (Methods 'Fish biomass and function'). Reef crests adjacent to rat-free islands are fully grazed nine times a year, compared to 2.8 times for rat-infested islands (median values; Fig. 4b; $\text{grazing}_{\text{rats}} - \text{grazing}_{\text{no rats}} = -1.18 [-2.24, -0.11]$, net rat effect). Although variable, median bioerosion rates were 94 tonnes $\text{ha}^{-1} \text{yr}^{-1}$ adjacent to rat-free islands, 3.8 times higher than the 24.5 tonnes $\text{ha}^{-1} \text{yr}^{-1}$ adjacent to rat-infested islands (Fig. 4c; $\text{erosion}_{\text{rats}} - \text{erosion}_{\text{no rats}} = -1.06 [-2.77, 0.53]$, net rat effect). Bioerosion is critical for breaking down dead reef corals between major disturbance events to provide stable substratum for new coral growth and recovery²⁰, and for providing sand to maintain island growth in low lying atolls²². While some bioeroding parrotfishes can take bites from corals, coral cover was not lower on rat-free islands (coral cover rat-free = $26.3\% \pm 5.2$ (mean \pm s.e.m.); rat-infested = $28.2\% \pm 5.5$). These data are consistent with seabirds on rat-free islands enhancing key ecosystem functions on coral reefs.

Following our surveys, coral reefs of the Chagos Archipelago lost approximately 75% coral cover in the 2016 El Niño-driven mass coral-bleaching event²³. It is possible that corals surrounding rat-free islands will show greater resilience to this event than corals adjacent to rat-infested islands, for two key reasons. First, in contrast to nutrient inputs from anthropogenic sources, nutrient delivery from biological sources, such as fish and seabirds, is rich in phosphorus^{3,24} and this has been shown to enhance coral thermo-tolerance²⁵ and coral calcification rates²⁴. Second, greater grazing rates, as observed on reefs adjacent to rat-free islands, is a key determinant of reef recovery²⁶.

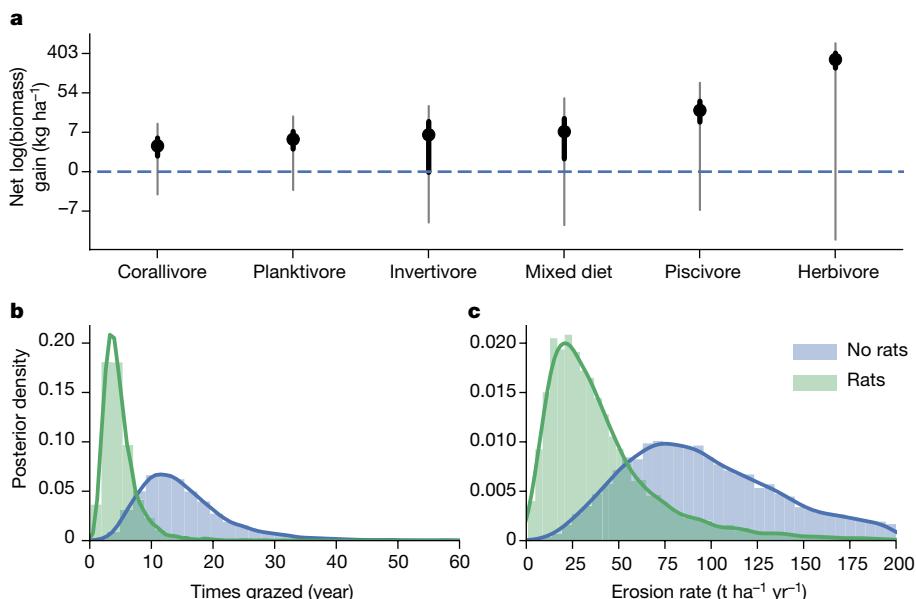


Fig. 4 | Biomass and functioning of reef-fish communities adjacent to islands with and without rats. **a**, Effect-size plots from a hierarchical Bayesian analysis of fish biomass for different feeding groups between rat-free and rat-infested islands. $n = 24$ biologically independent surveys for rat-free and rat-infested islands. Circles represent means and black and grey bars represent 50% and 95% uncertainty intervals, respectively

(highest posterior density). Positive values correspond to greater biomass on rat-free islands. **b**, Effect-size posterior-density distributions for the proportion of reef grazed by parrotfishes each year on rat-free versus rat-infested islands. **c**, Effect-size posterior density distributions for the volume of reef carbonate removed by parrotfishes each year on rat-free versus rat-infested islands.

Here, we show that seabird nutrient subsidies stimulate coral reef ecosystems, reflecting natural productivity and functioning in the absence of introduced rats. Oceanic coral reefs, such as those in the Chagos Archipelago, are highly productive ecosystems in an oligotrophic environment, the mechanisms of which have intrigued scientists for decades²⁷. Seabird-vectored nutrient subsidies are clearly a major pathway through which this productivity is supported, and such subsidies should be considered in the design and analyses of coral reef surveys adjacent to oceanic islands.

Rat eradication has been successful on 580 islands worldwide, and although success rates are slightly lower for tropical islands (89%) compared to temperate (96.5%), new techniques and guidelines are expected to close this gap²⁸. As eradication of rats from islands can lead to immigration and positive growth rates of seabird populations²⁹, rat removal should be a conservation priority for coral reef islands. The return of seabirds would benefit not only the island ecosystem, but also adjacent nearshore marine ecosystems. In a time of unprecedented threats to coral reefs from climate change³⁰, enhancing productivity and key ecosystem functions will give reefs the best possible chance to resist and recover from future disturbances.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0202-3>.

Received: 16 January 2018; Accepted: 9 May 2018;

Published online 11 July 2018.

1. Polis, G. A., Anderson, W. B. & Holt, R. D. Toward an integration of landscape and food web ecology: the dynamics of spatially subsidized food webs. *Annu. Rev. Ecol. Syst.* **28**, 289–316 (1997).
2. Doughty, C. E. et al. Global nutrient transport in a world of giants. *Proc. Natl. Acad. Sci. USA* **113**, 868–873 (2016).
3. Croll, D. A., Maron, J. L., Estes, J. A., Danner, E. M. & Byrd, G. V. Introduced predators transform subarctic islands from grassland to tundra. *Science* **307**, 1959–1961 (2005).
4. Fukami, T. et al. Above- and below-ground impacts of introduced predators in seabird-dominated island ecosystems. *Ecol. Lett.* **9**, 1299–1307 (2006).
5. Bump, J. K., Tischler, K. B., Schrank, A. J., Peterson, R. O. & Vucetich, J. A. Large herbivores and aquatic-terrestrial links in southern boreal forests. *J. Anim. Ecol.* **78**, 338–345 (2009).
6. Hocking, M. D. & Reynolds, J. D. Impacts of salmon on riparian plant diversity. *Science* **331**, 1609–1612 (2011).
7. Bouchard, S. S. & Bjorndal, K. A. Sea turtles as biological transporters of nutrients and energy from marine to terrestrial ecosystems. *Ecology* **81**, 2305–2313 (2000).
8. Jones, H. P. et al. Severity of the effects of invasive rats on seabirds: a global review. *Conserv. Biol.* **22**, 16–26 (2008).
9. Otero, X. L., De La Peña-Lastra, S., Pérez-Alberti, A., Ferreira, T. O. & Huerta-Díaz, M. A. Seabird colonies as important global drivers in the nitrogen and phosphorus cycles. *Nat. Commun.* **9**, 246 (2018).
10. Polis, G. A. & Hurd, S. D. Linking marine and terrestrial food webs: allochthonous input from the ocean supports high secondary productivity on small islands and coastal land communities. *Am. Nat.* **147**, 396–423 (1996).
11. McCauley, D. J. et al. From wing to wing: the persistence of long ecological interaction chains in less-disturbed ecosystems. *Sci. Rep.* **2**, 409 (2012).
12. Shatova, O., Wing, S. R., Gault-Ringold, M., Wing, L. & Hoffmann, L. J. Seabird guano enhances phytoplankton production in the Southern Ocean. *J. Exp. Mar. Biol. Ecol.* **483**, 74–87 (2016).
13. MacNeil, M. A. et al. Recovery potential of the world's coral reef fishes. *Nature* **520**, 341–344 (2015).
14. Carr, P. in *Important Bird Areas in the United Kingdom Overseas Territories* (ed. Sanders, S. M.) 37–55 (Royal Society for the Protection of Birds, Sandy, 2006).
15. Young, H. S., McCauley, D. J., Dunbar, R. B. & Dirzo, R. Plants cause ecosystem nutrient depletion via the interruption of bird-derived spatial subsidies. *Proc. Natl. Acad. Sci. USA* **107**, 2072–2077 (2010).
16. Szpak, P., Longstaffe, F. J., Millaire, J.-F. & White, C. D. Stable isotope biogeochemistry of seabird guano fertilization: results from growth chamber studies with maize (*Zea mays*). *PLoS ONE* **7**, e33741 (2012).
17. Lorrain, A. et al. Seabirds supply nitrogen to reef-building corals on remote Pacific islets. *Sci. Rep.* **7**, 3721 (2017).
18. McMahon, K. W., Johnson, B. J., Ambrose, W. G. Ocean Ecogeochemistry: a review. *Oceanogr. Mar. Biol. Annu. Rev.* **51**, 327–374 (2013).
19. Mora, C. *Ecology of Fishes on Coral Reefs* (Cambridge Univ. Press, Cambridge, 2015).
20. Bellwood, D. R., Hughes, T. P., Folke, C. & Nyström, M. Confronting the coral reef crisis. *Nature* **429**, 827–833 (2004).
21. Hoey, A. S. & Bellwood, D. R. Cross-shelf variation in the role of parrotfishes on the Great Barrier Reef. *Coral Reefs* **27**, 37–47 (2008).
22. Perry, C. T., Kench, P. S., O'Leary, M. J., Morgan, K. M. & Januchowski-Hartley, F. Linking reef ecology to island building: parrotfish identified as major producers of island-building sediment in the Maldives. *Geology* **43**, 503–506 (2015).
23. Sheppard, C. R. C. et al. Coral bleaching and mortality in the Chagos Archipelago. *Atoll Res. Bull.* **613**, 1–26 (2017).
24. Shantz, A. A. & Burkepile, D. E. Context-dependent effects of nutrient loading on the coral-algal mutualism. *Ecology* **95**, 1995–2005 (2014).
25. D'Angelo, C. & Wiedenmann, J. Impacts of nutrient enrichment on coral reefs: new perspectives and implications for coastal management and reef survival. *Curr. Opin. Environ. Sustain.* **7**, 82–93 (2014).
26. Graham, N. A. J., Jennings, S., MacNeil, M. A., Mouillot, D. & Wilson, S. K. Predicting climate-driven regime shifts versus rebound potential in coral reefs. *Nature* **518**, 94–97 (2015).
27. Gove, J. M. et al. Near-island biological hotspots in barren ocean basins. *Nature Commun.* **7**, 10581 (2016).
28. Keitt, B. et al. Best practice guidelines for rat eradication on tropical islands. *Biol. Conserv.* **185**, 17–26 (2015).
29. Brooke, M. de L. et al. Seabird population changes following mammal eradications on islands. *Anim. Conserv.* **21**, 3–12 (2018).
30. Hughes, T. P. et al. Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science* **359**, 80–83 (2018).

Acknowledgements This research was supported by the Australian Research Council's Centre of Excellence Program (CE140100020), a Royal Society University Research Fellowship awarded to N.A.J.G. (UF140691), and a Tier II NSERC Canada Research Chair awarded to M.A.M. We thank the British Indian Ocean Territory section of the British Foreign and Commonwealth Office for permission to conduct the study, and J. Turner for organizing the expedition. Animal ethics for fish collection were approved by James Cook University (approval number A2166). Thanks to J. Lokrantz for graphics help with Figs. 1, 2, and J. Barlow, S. Keith, and R. Evans for comments on the manuscript.

Reviewer information *Nature* thanks Y. Cherel, N. Knowlton and S. Wing for their contribution to the peer review of this work.

Author contributions N.A.J.G. conceived the study with S.K.W.; N.A.J.G., S.K.W. and P.C. collected the data; N.A.J.G., M.A.M., S.J. and A.S.H. developed and implemented the analyses; N.A.J.G. led the writing of the manuscript with S.K.W., M.A.M., S.J., A.S.H. and P.C.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0202-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to N.A.J.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Study sites. The Chagos Archipelago (British Indian Ocean Territory) is situated in the central Indian Ocean, due south of the Maldives (5° 50' S, 72° 00' E). The archipelago was first discovered in the early 16th century, but was not settled until the 18th century, after which rats were inadvertently introduced to some islands of the territory³¹. In the early 1970s the British government established a lease of the southernmost atoll (Diego Garcia) to the US Navy for a military base, and resettled the Chagossian people in Mauritius, Seychelles, and the UK. Since that time, the atolls of the northern archipelago have had very few direct human influences³², with exceptionally high reef-fish biomass³³, very low levels of water pollution³⁴, and there are currently ten designated (two more proposed) Important Bird and Biodiversity Areas¹⁴. However, invasive rats remain on a number of islands, creating a natural experiment to study the influence of rats on relatively undisturbed oceanic islands. In March–April 2015, we conducted research at 12 islands, across three atolls (Extended Data Table 2). Six of the islands were chosen as they are rat-free, whereas the other six are rat-infested.

Seabird surveys. Breeding-seabird densities on each island were counted annually from 2009–2015 using the apparently occupied nests methodology (AONs)^{35,36}. The entire coastline of each island was surveyed first and AONs were counted directly. Following the coastal survey, the interior of the island was searched. There were no breeding seabirds in the interior of Ile Poule, Grande Ile Mapou, Ile Fouquet, Eagle Island, and both Ile Anglaise islands. On islands for which the interior search revealed breeding seabirds, techniques to estimate AONs varied by family. Brown Booby (*S. leucogaster*) AONs were directly counted. Red-footed Booby (*S. sula*) AONs were counted directly except on Nelson's Island, and Grande Ile Coquillage. On these islands the total surface area of the breeding population was calculated using a handheld Global Positioning System (GPS) and random plots of the area were counted for AONs. Plot size and whereabouts was directed by accessibility, visibility and the vulnerability of the breeding population. The means of the AONs of the plots were multiplied by the number of plots possible in the mapped area to estimate total AONs. Breeding frigatebirds (*Fregata* sp.) on Nelson's Island and Grande Ile Coquillage were estimated using the same technique described above for the red-footed booby. No tropicbird (*P. lepturus*) nesting cavities were located, so tropicbird breeding numbers were estimated by counting aerial displaying pairs above islands or nest-prospecting adults in an appropriate habitat. Shearwaters (Procellariidae) nest in burrows, with the largest colony on South Brother island and some also on Nelson's Island. Burrows is a loose term that covers rock fissures, crevices, tree roots, coconut boles, and various underground holes. Island surface area with burrows was estimated and AONs estimated by multiplying from an average burrow density, taken from random 10 m² sample plots throughout the island. Burrows were assumed occupied when a bird or egg was seen in them, there were indications of use (for example, feathers, droppings), or they were heavily scented with shearwater musk. Burrows were assigned to either one of the shearwater species by identification of large chicks, eggs, or adults. For the species of arboreal-breeding noddy (*Anous* spp.), direct counts were impractical for the large colonies on South Brother and Nelson's Island, for which subsampling and multiplication to the total colony area was used. All ground-nesting tern species (Sternidae) with the exception of sooty tern (*O. fuscatus*) had AONs directly counted. To calculate the number of sooty tern AONs, the total colony area was mapped and random sample plots were counted for AONs to multiply up to the total area. Plot size was dictated by accessibility, visibility and to avoid disturbing dense aggregations of breeding birds, with numbers counted from outside the colony at random points around the perimeter. Although vegetation type, such as coconut versus native forests, can also affect bird densities³⁵, much of the indigenous island vegetation has been lost in Chagos¹⁴, and we used absolute bird-count estimates per island for this study.

Total annual seabird abundance was calculated on the basis of number of AONs multiplied by the mean number of birds occupying those nests per species, and the period of the year that the birds are present on the islands. For most species, a conservative estimate of three birds per nest was used (two adults and one chick), but some, for example sooty terns (*O. fuscatus*) have one adult, or one adult and one chick present for periods of the year, and others (for example, red-footed booby (*S. sula*)) have a chick and one to two juvenile/immature birds present in the nest or sub-colony area. The period of year spent on the island varied by species, from year round for species such as the brown booby (*S. leucogaster*) and common white tern (*G. alba*), to 4 months for the roseate tern (*Sterna dougallii*). Biomass of bird species was estimated using the average mass of an individual of each species taken from the Handbook of the Birds of the World³⁷.

We estimated the total nitrogen input from guano per hectare per year of each island following previously published methods¹⁵:

$$NI_{ij} = \frac{N_g \times Dr_i \times Bd_{ij} \times Res_{ij}}{\text{IsArea}_j}$$

where nitrogen input per hectare per year (NI) is estimated from the nitrogen content of guano (N_g), the defecation rate in g per species of bird (i) per day (Dr), the number of that species of bird (Bd) on the island (j), the number of days of the year that the species is resident on the island (Res), and the area of the island (IsArea). Nitrogen content of guano was held at 18.1% on the basis of guano samples from similar species in the Pacific¹⁵. The contribution of guano was based on the red-footed booby and scaled for other species on the basis of species biomass, assuming allometric relationships with body size¹⁵. We adjusted the Bd estimates to account for time off islands during feeding forays. Given uncertainties in foraging durations and whether birds would have full crops and bowels, it is hard to be completely precise in these calculations. We assigned the 14 species into three groups, which account for foraging excursions off island in a fairly conservative way.

Group 1: Tropical shearwater, wedge-tailed shearwater, white-tailed tropicbird, sooty tern, brown noddy, and frigatebirds. Foraging will vary during the breeding cycle, but often one adult is foraging and may be off the island overnight. We therefore assumed only one adult of the pair was on the island at any one time.

Group 2: Red-footed booby. One bird of the pair makes daylight foraging forays but returns overnight. Adult numbers were therefore halved only during daylight hours (12 h).

Group 3: Great crested tern, roseate tern, black-naped tern, common white tern, bridled tern, brown booby, lesser noddy. In Chagos, these species tend to make much shorter foraging forays (1–4 h depending on species), meaning defecation at sea will be minimal compared to land. We therefore did not make any adjustments to their numbers.

Seabird densities per hectare of rat-free versus rat-infested islands were plotted as notched box plots, in which the horizontal line is the median, box height depicts the interquartile range, and diagonal notches in the boxes illustrate the 95% confidence interval around the median³⁸. The biomass of families of birds per island were plotted as log-scale heat maps for rat-free and rat-infested islands. Nitrogen input for rat-free versus rat-infested islands was plotted as notched box plots of kg per hectare per year. We also developed a set of simple Bayesian models to estimate the net rat effects on log-scale bird numbers, log-scale total biomass, and nitrogen input between rat-free and rat-infested islands:

$$y_i \sim N(\mu_i, \sigma_i)$$

$$\mu_i = \beta_0 + \beta_1 \times \text{RAT}$$

$$\beta_{0,1} \sim N(0, 10)$$

$$\sigma_i \sim U(0, 10)$$

where y_i was the response variable, RAT was a dummy variable for rat-infested islands, and variances were estimated independently within treatments. The β_1 parameters are the rat effect sizes reported in the caption of Fig. 1 along with the proportion of β_1 posterior density below zero.

Primary production and potential prey biomass and production available to seabirds. Biomass, production, and size structure of consumers in the ocean surrounding the Chagos Islands were calculated from the primary production available to support them using a size-based model that characterizes some of the main factors affecting the rate and efficiency of energy processing in marine ecosystems³⁹. In brief, these factors are (i) temperature, which affects rates of metabolism and hence growth and mortality; (ii) the size of phytoplankton and the predator to prey body mass ratio, which determine the number of steps in a food chain; and (iii) trophic transfer efficiency, a measure of the energy conserved and lost at each step in the chain. In the model, size composition of the phytoplankton community is predicted from primary production and temperature using empirical relationships and, in turn, this size composition is used to estimate particle export ratios that influence transfer efficiency in the first steps of the food chain. The model is depth integrated and we made the simplifying assumption that all primary production occurs in the euphotic zone. We did not explicitly model production of benthic communities, but these would not be accessible to seabirds.

In the model, relationships between primary-consumer production and consumer production at any higher trophic level are determined by trophic transfer efficiency. Production at a given body mass or trophic level was converted to biomass and numbers at the same body mass or trophic level on the basis of the assumption that body size and temperature determined individual rates of production³⁹. The modelled size spectrum was discretized into units of 0.1 (log₁₀) for analysis.

The environmental data used to force the models consisted of annual mean estimates of depth-integrated primary production (g C m⁻² d⁻¹) and sea surface temperature (°C) as derived from monthly predictions for the years 2010–2012.

Chlorophyll and primary production were obtained from the Mercator Ocean Project (Global Biogeochemical Analysis Product, BIOMER1V1 monthly 0.5° degree resolution) (<http://www.mercator-ocean.fr/>; the data are copyright of Mercator Ocean, product and interpretations obtained from Mercator Ocean products, Mercator Ocean cannot be held responsible for the results nor for the use to which they are put, all rights reserved.). Monthly temperature data were obtained from the Mercator Ocean physical NEMO model (PSY3V3R1)⁴⁰. Inputs to the size-based models were allocated to a 0.5° grid that covered the sea area defined by the maximum foraging distance of the species of seabird, assumed to be a radius, such that foraging areas were circular around the islands (Extended Data Table 1). These distances are an approximation from the published literature, given that foraging ranges can vary geographically⁴¹. Cells were assigned a mixed layer depth (m) and total depth (m)^{42,43}. Mean biomass and production for organisms in body mass (wet weight) classes 0.1–9 g (smaller prey) and 1–50 g (larger prey) was estimated per unit area by grid cell, to approximate size ranges consumed by the seabird species on the basis of prey-size information in the literature and body-mass class^{44,45}. To address considerable uncertainty in model parameters, we ran 10,000 simulations for each biomass or production estimate in each grid cell, with parameter estimates in each simulation drawn randomly from appropriate distributions. When parameters were correlated, the parameter estimates were drawn from multivariate distributions³⁹. Model results, expressed as medians and percentiles, were calculated from the distribution of output values. Conversions from carbon to wet weight were based on published values^{39,46}. Estimates of nitrogen content in prey-size classes were based on an assumed C:N ratio of 3.4:1, which is a typical value for fish⁴⁷ and reflects the fall in the C:N ratio with trophic level in food webs that are supported by primary producers with C:N ratios typically averaging 6.6:1^{48,49}. Estimates of biomass and production per unit area were converted to estimates of total biomass or production in the foraging area of each bird species (Extended Data Table 1 and Extended Data Fig. 1).

While rates of gross primary production can be high on coral reefs, net primary production, although variable in space and time, is typically comparable with net primary production in the more productive areas of the tropical ocean^{50,51}. Given the area of reef surrounding the rat-free islands is approximately 1.02 km², whereas foraging areas are >105 km² for 14 of the 15 bird species using these islands, large numbers of seabirds can feed from oceanic food webs with much higher production than those on the reefs (Extended Data Fig. 1). Even the production estimates for prey in the size ranges eaten by the seabirds are typically three or more orders of magnitude higher than the expected primary production on this area of reef (0.0001 Tg C yr⁻¹, if mean on-reef primary production is assumed to be 0.3 g C m⁻² d⁻¹)⁵⁰. Given the numbers of seabirds and the extent of the prey resource they have the potential to access, the strong signal from guano-derived nitrogen on the reefs surrounding rat-free islands is unsurprising. While the model has a number of assumptions, the results do highlight that oceanic production in the foraging area is expected to be several orders of magnitude higher than production on the reefs surrounding the islands and therefore that the higher levels of connectivity that result from higher seabird abundance have the potential to transport relatively high quantities of nitrogen to the reef systems.

Isotope sampling. From each island, ten samples of topsoil (<5 cm from surface) were taken from just behind the coastal vegetation boundary. Loose leaf litter and other vegetation was cleared to expose the soil, and samples were taken a minimum of 10 m apart. Along the beach margin of each island, new-growth leaf samples were taken from ten *S. taccada* plants. On the reef flat on the lagoonal side of each island (1 m deep and approximately 100 m from shore) ten samples of filter-feeding sponges (*Spheciopspongia* sp.) and macroalgae (*Halimeda* sp.) were taken from individual colonies and thalli, respectively. On the reef crest of each island (~3 m deep and 230 ± 55 m from shore) ten turf-algal samples were taken from dead corals. Ten adult territorial herbivorous damselfish (*P. lacrymatus*) individuals were collected on the reef crest of each island in the same area the turf algae were collected. Fish were euthanized on ice. Fish samples could not be collected from Nelson's Island. A sample of dorsal white muscle was taken from each fish. All samples were dried in a drying oven at 60 °C for 24 h or until fully dry. Samples were powdered with a pestle and mortar and stored in sealed plastic sample vials.

Stable isotope analysis of nitrogen for all samples was carried out at the University of Windsor, Canada. Isotope ratios were calculated from 400 to 600 µg of each sample added to tin capsules and analysed with a continuous-flow isotope-ratio mass spectrometer (Finnigan MAT Deltaplus, Thermo Finnigan). Total nitrogen content (%) was also estimated. Stable isotope values for nitrogen are expressed as delta (δ) values for the ratio of ¹⁵N:¹⁴N. Turf, sponge, soil and macroalgae samples were acid washed with hydrochloric acid to dissolve any calcareous matter or sediments that may have contaminated the samples. Subsets of samples that were run with and without the acid wash had correlation coefficients between 0.9 (turf-algae) and 0.99 (soil), and all samples from rat-free and rat-infested islands were treated the same. The standard reference material was atmospheric nitrogen. Samples were run twice, with select samples run in triplicate

to ensure accuracy of readings. Accuracy was within 0.3‰ for soil and within 0.1‰ for other samples, on the basis of soil elemental microanalysis B2153 and USGS 40 internal standards, respectively.

¹⁵N values between rat-free and rat-infested island treatments were analysed using Bayesian hierarchical models, with the area of reef surrounding each island (RA; calculated using GIS) as a covariate, and samples nested within their specific atoll. Distance to shore from the reef crest (DS) was used as an additional covariate for the turf algae and fish muscle samples. Models were run using the PyMC3 package⁵² in Python (www.python.org), including a *t*-distribution with four degrees of freedom as:

$$\delta^{15}\text{N}_{oij} \sim t_4(\mu_{oij}, \sigma_0)$$

$$\mu_{oij} = \beta_{0i} + \beta_{1o} + \beta_{2o}\text{RA}_j + \beta_3\text{DS}_j$$

$$\beta_{0i} \sim N(\gamma_0, \sigma_\gamma)$$

$$\beta_{1,2,3}, \gamma_0 \sim N(0, 1,000)$$

$$\sigma_0, \sigma_\gamma \sim U(0, 100)$$

where each organism (*o*) had their own offset (β_{0i}) relative to island-level (*i*) soil intercepts (β_1). Models were examined for convergence and fit by consideration of stability in posterior chains, Gelman–Rubin (\bar{R}) statistics, and the fit of the models with the data⁵³.

Fish growth. The total length of each damselfish (*P. lacrymatus*) sampled was carefully measured to the nearest mm. The paired sagittal otoliths (ear bones) were removed from each individual to estimate age⁵⁴. One otolith from each pair was weighed to the nearest 0.0001 g and affixed to a glass slide using thermoplastic glue with the primordium located just inside the edge of the slide and the sulcus ridge perpendicular to the slide edge. The otolith was ground to the slide edge using a 600-grit diamond lapping disc on a grinding wheel along the longitudinal axis. The otolith was then removed and re-affixed to a clean slide with the flat surface against the slide face and ground to produce a thin transverse section approximately 200 µm thick, encompassing the core material. Finally, the exposed section was covered in thermoplastic glue to improve clarity of microstructures. Sections were examined twice and age in years was estimated by counting annuli (alternating translucent and opaque bands) along a consistent axis on the ventral side of the sulcus ridge, using transmitted light on a stereo microscope.

Growth curves for the otoliths from the rat-free versus rat-infested islands were modelled using the three-parameter van Bertalanffy growth function, implemented in PyMC3 as:

$$\begin{aligned} \log(L_{t,i}) &\sim N(\mu_p, \sigma_0) \\ \mu_i &= \log(L_\infty - (L_\infty - L_0)e^{-(k_0 + k_1)t_i}) \\ k_0 &\sim U(0.001, 1) \\ k_1 &\sim N(0, 10) \\ L_0 &\sim N(0, \min(L_t)) \\ L_\infty &\sim U(\max(L_t), \max(L_t) \times 2) \\ \sigma_0 &\sim U(0, 1,000) \end{aligned}$$

Where L_t is the observed total length (cm) at age t (years), L_∞ is the estimated asymptotic length, K is the coefficient used to describe the curvature of growth towards L_∞ (here split into k_0 (no rats) and k_1 (rat offset)) and L_0 is the theoretical length at age zero⁵⁵. We specified uniform bounds for the L parameters on the basis of observed minimum ($\min(L) = 6.2$) and maximum ($\max(L) = 10.4$) fish lengths. Again, models were examined for convergence and fit by consideration of stability in posterior chains, \bar{R} statistics, and the fit of the models with the data.

Fish biomass and function. Underwater visual surveys were conducted along the reef crest of each island on the lagoonal side of each atoll. Four 30-m transects were laid along the reef crest at 3 m depth, separated by at least 10 m. Benthic cover of corals, algae, and other organisms were surveyed using the line intercept method, for which the substratum type under the transect tape was recorded along the entire 30-m length. The structural complexity of the reef was estimated visually on a six-point scale, ranging from no relief to exceptionally complex (>1 m high) relief with numerous caves and overhangs. This structural complexity measure captures landscape complexity, including the complexity provided by live corals, that of the underlying reef matrix and other geological features, and has been shown to correlate well to other measures of complexity, such as measures of reef height and the linear versus contour chain method⁵⁶. The density and individual sizes of

diurnally active, non-cryptic species of reef-associated fish were estimated along each transect. Larger, more active fish were surveyed on the first pass of each transect in a 5-m-wide belt, whereas the more territorial and abundant damselfish family (Pomacentridae) were surveyed on a second pass of the transect in a 2-m-wide belt. We converted data on fish counts to biomass with published length-weight relationships from FishBase (<http://www.fishbase.org>) and a previously published work⁵⁷. Fish were assigned to feeding groups on the basis of their dominant diets and feeding behaviour⁵⁸.

The grazing and erosion potential (that is, area of reef scraped and volume of carbonates removed, respectively) by parrotfishes at each site was calculated as the product of feeding rate, bite dimension (area or volume), and fish density (following previously published methods²¹). Size-specific feeding rates for each species were derived from best-fit regressions of bite rate (bites per min) and fish length (total length, cm) for each species. Bite rates were quantified at three locations (Lizard Island, northern Great Barrier Reef, northern Sumatra, Indonesia, and the central Red Sea) using focal feeding observations. An individual parrotfish was haphazardly selected, followed for a short period of acclimation (~1 min) during which the fish length (total length, TL) was estimated to the nearest centimetre. After the acclimation period each fish was followed for a minimum of 3 min during which the number of bites on different benthic substrata (primarily epilithic algal matrix and live corals) and observation time were recorded. Bite rates were then converted to bites per min. Observations were discontinued if the focal individual displayed a detectable response to the diver. All feeding observations were conducted from 9:00 to 15:00 with a minimum of 25 observations conducted per species per location.

The area (mm^2) and volume (mm^3) of material removed per bite by individual parrotfish was estimated from species-specific relationships between bite size and fish length. To estimate bite area an individual parrotfish was haphazardly selected, its total length was estimated and it was followed until it took a bite from the reef substratum. The dimensions of the bite (length and width) were then measured in situ using dial callipers. A minimum of 16 observations (mean = 34.3 observations) were made per species, with all observations performed at Lizard Island, northern GBR. Bite volumes of species were largely taken from the literature⁵⁹, and supplemented with in situ observations at Lizard Island for *Chlorurus microrhinos*. Where possible, species-specific bite rates and bite dimensions were used, when these were not available, values for closely related congeners were used.

Total biomass and biomass of each trophic feeding group of fish (BIO_{ij}) was modelled using Bayesian hierarchical models, with observations (j) nested within atolls (i) and including factors that could influence fish biomass as covariates; coral cover (HC) and reef structural complexity (SC). The general model was:

$$\log(\text{BIO}_{fij}) \sim N(\mu_{fj}, \sigma_0)$$

$$\mu_{fj} = \beta_{f0i} + \beta_1 \text{RAT} + \beta_2 \text{SC}_j + \beta_3 \text{HC}_j$$

$$\beta_{f0i} \sim N(\gamma_{f0}, \sigma_\gamma)$$

$$\beta_{1,2,3}, \gamma_{f0} \sim N(0, 1,000)$$

$$\sigma_0, \sigma_\gamma \sim U(0, 100)$$

with models examined for convergence and fit by consideration of stability in posterior chains, \hat{R} statistics and the fit of the models with the data.

The two ecosystem functions, grazing and erosion potential (rounded to nearest whole number), were modelled with the same Bayesian hierarchical structure, but with an alternative Poisson (Pois) rate (XR) likelihood:

$$\text{XR}_{ij} \sim \text{Pois}(e^{\mu_{ji}})$$

$$\mu_{ji} = \beta_{j0i} + \beta_1 \text{RAT} + \beta_2 \text{SC}_j + \beta_3 \text{HC}_j$$

$$\beta_{j0i} \sim N(\gamma_{j0}, \sigma_\gamma)$$

$$\beta_{1,2,3}, \gamma_{j0} \sim N(0, 1,000)$$

$$\sigma_0, \sigma_\gamma \sim U(0, 100)$$

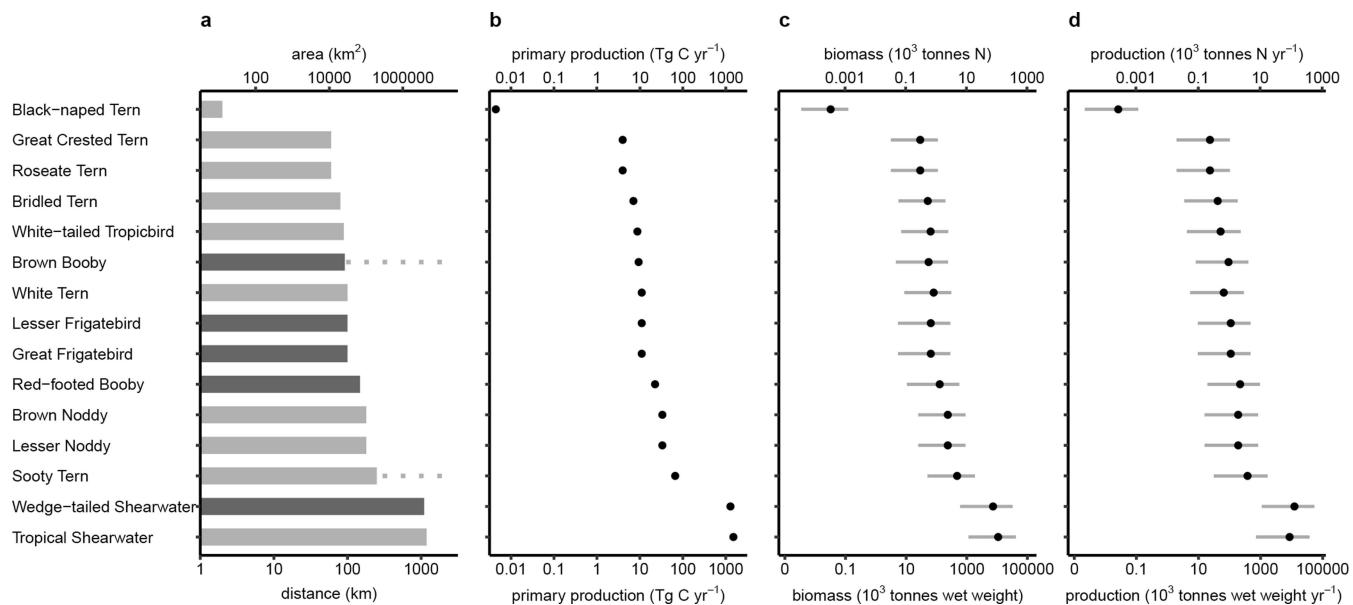
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Code used for Figs. 1–4 in this paper are available from GitHub (<https://github.com/mamacneil/ChagosRats>).

Data availability. Data used for Figs. 1–4 in this paper are available from GitHub (<https://github.com/mamacneil/ChagosRats>).

31. Wenban-Smith, N. & Carter, M. *Chagos: a History: Exploration, Exploitation, Expulsion* (Chagos Conservation Trust, London, 2016).
32. Sheppard, C. R. C. et al. Reefs and islands of the Chagos Archipelago, Indian Ocean: why it is the world's largest no-take marine protected area. *Aquat. Conserv.* **22**, 232–261 (2012).
33. Graham, N. A. J. et al. Human disruption of coral reef trophic structure. *Curr. Biol.* **27**, 231–236 (2017).
34. Readman, J. W. et al. in *Coral Reefs of the United Kingdom Overseas Territories* (ed. Sheppard, C. R. C.) 283–298 (Springer, Dordrecht, 2013).
35. Bibby, C. J., Burgess, N. B. & Hill, D. A. *Bird Census Techniques* (Academic, London, 1992).
36. McGowan, A., Broderick, A. C. & Godley, B. J. Seabird populations of the Chagos Archipelago: an evaluation of IBA sites. *Oryx* **42**, 424–429 (2008).
37. del Hoyo, J., Elliott, A., Sargatal, J., Christie, D. A. & de Juana, E. (eds) *Handbook of the Birds of the World Alive* (Lynx Edicions, Barcelona, 2017).
38. Krzywinski, M. & Altman, N. Visualizing samples with box plots. *Nat. Methods* **11**, 119–120 (2014).
39. Jennings, S. & Collingridge, K. Predicting consumer biomass, size-structure, production, catch potential, responses to fishing and associated uncertainties in the world's marine ecosystems. *PLoS ONE* **10**, e0133794 (2015).
40. Aumont, O. & Bopp, L. Globalizing results from ocean in situ iron fertilization studies. *Glob. Biogeochem. Cycles* **20**, GB2017 (2006).
41. Mendez, L. et al. Geographical variation in the foraging behaviour of the pantropical red-footed booby. *Mar. Ecol. Prog. Ser.* **568**, 217–230 (2017).
42. Schmidtko, S., Johnson, G. C. & Lyman, J. M. MIMOC: A global monthly isopycnal upper-ocean climatology with mixed layers. *J. Geophys. Res.* **118**, 1658–1672 (2013).
43. IOC, IHC, BODC. The GEBCO Digital Atlas. (BODC, 2008).
44. Ashmole, N. P. Body size, prey size, and ecological segregation in five sympatric tropical terns (Aves: Laridae). *Syst. Zool.* **17**, 292–304 (1968).
45. Harrison, C. S., Hida, T. S. & Seki, M. P. Hawaiian seabird feeding ecology. *Wildl. Monogr.* **85**, 3–71 (1983).
46. Wiebe, P. H., Boyd, S. H. & Cox, J. L. Relationships between zooplankton displacement volume, wet weight, dry weight and carbon. *Fish Bull.* **73**, 777–786 (1975).
47. Jennings, S. & Cogan, S. M. Nitrogen and carbon stable isotope variation in northeast Atlantic fishes and squids. *Ecology* **96**, 2568 (2015).
48. Martiny, A. C., Vrugt, J. A. & Lomas, M. W. Concentrations and ratios of particulate organic carbon, nitrogen, and phosphorus in the global ocean. *Sci. Data* **1**, 140048 (2014).
49. Martiny, A. C., Vrugt, J. A., Primeau, F. W. & Lomas, M. W. Regional variation in the particulate organic carbon to nitrogen ratio in the surface ocean. *Glob. Biogeochem. Cycles* **27**, 723–731 (2013).
50. Crossland, C. J., Hatcher, B. G. & Smith, S. V. Role of coral reefs in global ocean production. *Coral Reefs* **10**, 55–64 (1991).
51. Hatcher, B. G. Coral reef primary productivity. A hierarchy of pattern and process. *Trends Ecol. Evol.* **5**, 149–155 (1990).
52. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2**, e55 (2016).
53. Gelman, A. et al. *Bayesian Data Analysis* Vol. 2 (CRC, Boca Raton 2014).
54. Campana, S. E. Otolith science entering the 21st century. *Mar. Freshw. Res.* **56**, 485–495 (2005).
55. Pardo, S. A., Cooper, A. B. & Dulvy, N. K. Avoiding fishy growth curves. *Methods Ecol. Evol.* **4**, 353–360 (2013).
56. Wilson, S. K., Graham, N. A. J. & Polunin, N. V. C. Appraisal of visual assessments of habitat complexity and benthic composition on coral reefs. *Mar. Biol.* **151**, 1069–1076 (2007).
57. Letourneau, Y. Length-weight relationships of some marine fish species in Réunion Island, Indian Ocean. *Naga* **21**, 37–39 (1998).
58. Wilson, S. K. et al. Exploitation and habitat degradation as agents of change within coral reef fish communities. *Glob. Chang. Biol.* **14**, 2796–2809 (2008).
59. Ong, L. & Holland, K. N. Bioerosion of coral reefs by two Hawaiian parrotfishes: species, size differences and fishery implications. *Mar. Biol.* **157**, 1313–1323 (2010).
60. McDue, F., Weeks, S. J., Miller, M. G. R. & Congdon, B. C. Breeding tropical shearwaters use distant foraging sites when self-provisioning. *Mar. Ornithol.* **43**, 123–129 (2015).
61. Calabrese, L. *Foraging Ecology and Breeding Biology of Wedge-Tailed Shearwater (Puffinus pacificus) and Tropical Shearwater (Puffinus bailloni) on Aride Island Nature Reserve, Seychelles: Tools for Conservation*. PhD thesis, Université Pierre et Marie Curie-Paris VI (2015).
62. Pennycuick, C. J., Schaffner, F. C., Fuller, M. R., Obrecht, H. H. III & Sternberg, L. Foraging flights of the white-tailed tropicbird (*Phaethon lepturus*): radiotracking and doubly-labelled water. *Colon. Waterbirds* **13**, 96–102 (1990).
63. Jaquemet, S., Le Corre, M., Marsac, F., Potier, M. & Weimerskirch, H. Foraging habitats of the seabird community of Europa Island (Mozambique Channel). *Mar. Biol.* **147**, 573–582 (2005).

64. Gilardi, J. D. Sex-specific foraging distributions of brown boobies in the eastern tropical Pacific. *Colon. Waterbirds* **15**, 148–151 (1992).
65. Weimerskirch, H., Le Corre, M., Jaquemet, S. & Marsac, F. Foraging strategy of a tropical seabird, the red-footed booby, in a dynamic marine environment. *Mar. Ecol. Prog. Ser.* **288**, 251–261 (2005).
66. Surman, C. A. & Wooller, R. D. Comparative foraging ecology of five sympatric terns at a sub-tropical island in the eastern Indian Ocean. *J. Zool. (Lond.)* **259**, 219–230 (2003).
67. Bourne, W. R. P. & Simmons, K. E. L. The distribution and breeding success of seabirds on and around Ascension in the tropical Atlantic Ocean. *Atl. Seabirds* **3**, 187–202 (2001).
68. Dunlop, J. N. Foraging range, marine habitat and diet of bridled terns breeding in Western Australia. *Corella* **21**, 77–82 (1997).
69. Hulsmann, K. & Smith, G. Biology and growth of the black-naped tern (*Sterna sumatrana*): A hypothesis to explain the relative growth rates of inshore, offshore and pelagic feeders. *Emu* **88**, 234–242 (1988).



Extended Data Fig. 1 | Primary production and potential prey biomass and production in areas accessible to seabirds foraging around the Chagos Islands. **a**, Recorded foraging ranges for seabird species that feed on smaller prey (light tone, 0.1–9 g individual wet weight) or larger prey (dark tone, 1–50 g individual wet weight; broken lines indicate that greater ranges are expected for two of the species thus foraging area calculations assumed that the foraging range is the radius of the foraging

area). **b**, Primary production in the foraging area. **c**, Modelled biomass. **d**, Production of fauna in the foraging area. Median and 90% uncertainty intervals on the basis of 10,000 simulations to assess the effects of parameter uncertainty³⁹ on biomass or production estimates are shown. Biomass and production were estimated for fauna in the prey size ranges consumed by each bird species, and expressed as wet and nitrogen (N) weight, respectively.

Extended Data Table 1 | Species-specific foraging locations, foraging distances and foraging observations from Chagos

Species	Foraging location	Reported foraging distance in km	Chagos-specific foraging behaviour
Wedge-tailed Shearwater <i>Ardenna pacifica</i>	Pelagic	300-1100 (ref. 60)	Forages over open ocean normally well away from the islands. Often associates with other seabirds where sub-surface predators (e.g. tuna) are driving prey to the surface.
Tropical Shearwater <i>Puffinus bailloni</i>	Mainly pelagic	379-1190 (ref. 61)	As <i>Ardenna pacifica</i> .
White-tailed Tropicbird <i>Phaethon lepturus</i>		89 (ref. 62)	Solitary; forages over open ocean far away from land.
Lesser Frigatebird <i>Fregata ariel</i>	Pelagic	100 (ref. 63)	Normally pelagic, occasionally kleptoparasites other seabirds returning to colonies with food
Great Frigatebird <i>Fregata minor</i>	Pelagic	100 (ref. 64)	As <i>Fregata ariel</i>
Brown Booby <i>Sula leucogaster</i>	Inshore waters	>92 (ref. 64)	Usually seen foraging over submerged off-atoll banks. Presumably this is because there is very little "inshore waters" in the Chagos.
Red-footed Booby <i>Sula sula</i>	Pelagic	39-148 (ref. 65)	Forages over open ocean areas usually in association with tuna. At large prey balls, hundreds of red-footed boobys can be present, along with shearwaters, brown noddy, sooty terns and white terns.
Brown Noddy <i>Anous stolidus</i>	Pelagic	180 (ref. 66)	Forages over open ocean miles away from land. Often feeds in association with Red-footed Booby and on prey balls.
Lesser Noddy <i>Anous tenuirostris</i>	Pelagic	180 (ref. 66)	In the Chagos this species forages in flocks in atoll lagoons or over off-atoll banks. Usually associated with tuna.
White Tern <i>Gygis alba</i>	Semi-pelagic, including inshore waters	2-100 (ref. 67)	Normally found foraging within 5km of land. Targets large prey balls.
Sooty Tern <i>Onychoprion fuscata</i>	Pelagic	>250 (ref. 63)	Forages far and wide over open ocean.
Bridled Tern <i>Onychoprion anaethetus</i>	Semi-pelagic, including inshore waters	20-80 (ref. 68)	Usually seen foraging over off-atoll banks.
Roseate Tern <i>Sterna dougallii</i>	Semi-pelagic, including inshore waters	60 (ref. 66)	Very rare in the Chagos and its foraging habits are not known
Black-naped Tern <i>Sterna sumatrana</i>	Atoll lagoons, close inshore, but sometimes at sea.	2 (ref. 69)	Feeds around coral fringes of islands or in lagoons. Often associates with lesser noddy.
Great Crested Tern <i>Thalasseus bergii</i>	Atoll lagoons, close inshore, but sometimes at sea.	60 (ref. 66)	Another inshore species with similar foraging requirements as black-naped tern.

Data on forage distances are from previously published work⁶⁰⁻⁶⁹.

Extended Data Table 2 | Islands used in the study

Treatment	Atoll	Island	South	East	Island area (ha)
Rat-free	Peros Banhos	Ile Longue	5.270	71.867	25.5
Rat-free	Peros Banhos	Grande Ile Coquillage	5.372	71.969	28
Rat-free	Salomon	Ile de la Passe	5.304	72.251	26
Rat-free	Great Chagos Bank	Nelson's Island	5.682	72.313	81
Rat-free	Great Chagos Bank	Middle Brother	6.154	71.517	8
Rat-free	Great Chagos Bank	South Brother	6.172	71.544	23
Rat-infested	Peros Banhos	Ile Anglaise	5.439	71.757	12
Rat-infested	Peros Banhos	Ile Poule	5.414	71.755	108
Rat-infested	Peros Banhos	Grande Ile Mapou	5.266	71.753	19.5
Rat-infested	Salomon	Ile Fouquet	5.343	72.262	39.5
Rat-infested	Salomon	Ile Anglaise	5.327	72.223	75.5
Rat-infested	Great Chagos Bank	Eagle Island	6.187	71.338	243.5

Robust relationship between air quality and infant mortality in Africa

Sam Heft-Neal¹, Jennifer Burney², Eran Bendavid³ & Marshall Burke^{1,4,5*}

Poor air quality is thought to be an important mortality risk factor globally^{1–3}, but there is little direct evidence from the developing world on how mortality risk varies with changing exposure to ambient particulate matter. Current global estimates apply exposure–response relationships that have been derived mostly from wealthy, mid-latitude countries to spatial population data⁴, and these estimates remain unvalidated across large portions of the globe. Here we combine household survey-based information on the location and timing of nearly 1 million births across sub-Saharan Africa with satellite-based estimates⁵ of exposure to ambient respirable particulate matter with an aerodynamic diameter less than 2.5 μm ($\text{PM}_{2.5}$) to estimate the impact of air quality on mortality rates among infants in Africa. We find that a $10 \mu\text{g m}^{-3}$ increase in $\text{PM}_{2.5}$ concentration is associated with a 9% (95% confidence interval, 4–14%) rise in infant mortality across the dataset. This effect has not declined over the last 15 years and does not diminish with higher levels of household wealth. Our estimates suggest that $\text{PM}_{2.5}$ concentrations above minimum exposure levels were responsible for 22% (95% confidence interval, 9–35%) of infant deaths in our 30 study countries and led to 449,000 (95% confidence interval, 194,000–709,000) additional deaths of infants in 2015, an estimate that is more than three times higher than existing estimates that attribute death of infants to poor air quality for these countries^{2,6}. Upward revision of disease-burden estimates in the studied countries in Africa alone would result in a doubling of current estimates of global deaths of infants that are associated with air pollution, and modest reductions in African $\text{PM}_{2.5}$ exposures are

predicted to have health benefits to infants that are larger than most known health interventions.

Epidemiological studies consistently highlight poor air quality as an important contributor to death and disability, with recent estimates showing that exposure to ambient $\text{PM}_{2.5}$ is associated with 3–4 million global deaths annually^{1,2}. Such estimates are influential in a wide variety of research activities^{3,7,8} and policy decisions, including the allocation of health resources, the designation of pollution standards, and the adoption of climate-mitigation policies.

However, the relationship between air quality and mortality in the developing world—where a large proportion of the deaths that can be attributable to poor air quality are thought to occur—remains poorly quantified, limiting our understanding of relative disease burdens and appropriate policy responses. Broad-scale evidence on the health burden of exposure to ambient air pollution comes mainly from developed countries^{1,2,4}, where co-morbidities differ and where both mortality rates and average ambient $\text{PM}_{2.5}$ concentrations are typically much lower (Extended Data Fig. 1). In much of the developing world, limited air pollution data make quantification of dose–response functions challenging^{9,10}, and it is unknown whether large recent declines in infant mortality¹¹ would increase the relative health effects of poor air quality (if other unrelated causes of death are now less important) or decrease them (if children are more resilient).

Here we quantify associations between air quality and the health of infants in Africa by combining recent satellite-based measurements of annual ambient $\text{PM}_{2.5}$ concentrations with household survey data on

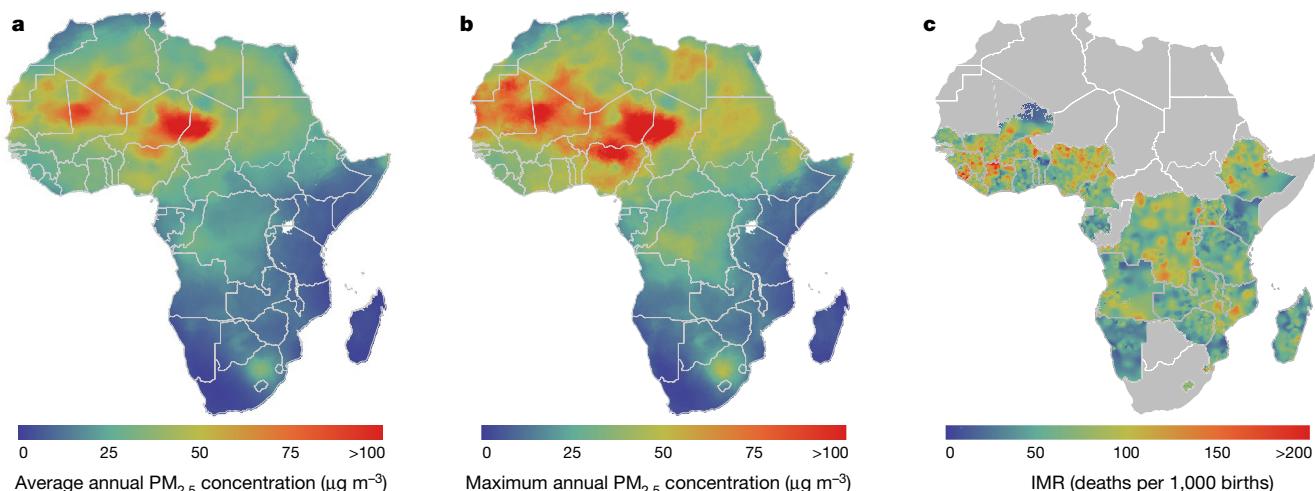


Fig. 1 | Spatial patterns of pollution and infant mortality in Africa.

a, Long-run average $\text{PM}_{2.5}$ concentration for 2001–2015⁵. **b**, Maximum annual $\text{PM}_{2.5}$ concentration for 2001–2015. **c**, Average infant mortality rate

(IMR) in study countries for 2001–2015, derived from Demographic and Health Surveys using previously described methods¹¹. Country outlines were obtained from Global Administrative Areas, version 2.0³⁰.

¹Center on Food Security and the Environment, Stanford University, Stanford, CA, USA. ²School of Global Policy and Strategy, University of California, San Diego, San Diego, CA, USA. ³School of Medicine, Stanford University, Stanford, CA, USA. ⁴Department of Earth System Science, Stanford University, Stanford, CA, USA. ⁵National Bureau of Economic Research, Cambridge, MA, USA. *e-mail: mburke@stanford.edu

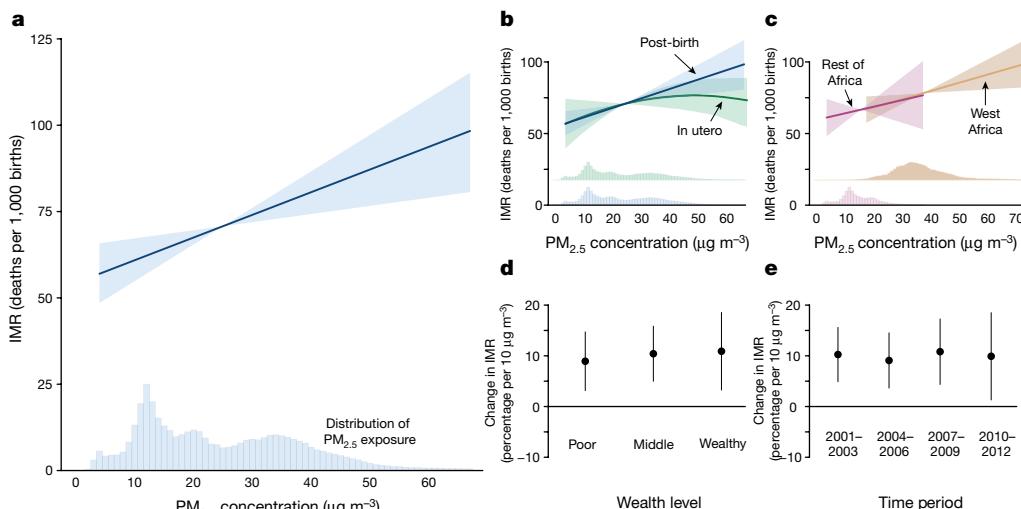


Fig. 2 | Mortality of infants in Africa is strongly and linearly increasing with post-birth PM_{2.5} exposure. **a**, Effect of PM_{2.5} exposure during the 12 months after birth on mortality rates of infants ($n = 990,696$ births). Response function is centred at mean PM_{2.5} concentration ($25 \mu\text{g m}^{-3}$) and mean IMR (71 deaths per 1,000 births). Histogram shows the distribution of exposures across sample locations. **b**, Impacts of in utero versus post-birth exposures. **c**, Impacts of post-birth exposure in West Africa (higher exposure) versus the rest of Africa (lower exposure). See Extended Data Fig. 2b for countries in each region. **d**, Effect of post-birth exposure on child mortality by terciles of household-level asset wealth, measured as the percentage change in infant mortality per $10 \mu\text{g m}^{-3}$ increase in PM_{2.5} exposure. **e**, Effect of post-birth PM_{2.5} exposure on IMR over time, measured as the percentage change in IMR per $10 \mu\text{g m}^{-3}$ increase in PM_{2.5} exposure.

infant mortality (death in the first 12 months of life) as measured in the Demographic and Health Surveys, a set of nationally representative household health surveys. We use data from 65 available Demographic and Health Surveys across 30 sub-Saharan African (SSA) countries carried out between 2001 and 2015, representing 990,696 births over the period (Extended Data Fig. 2). We match the location and timing of each birth to satellite-based estimates of PM_{2.5} exposure from 9 months before to 12 months after birth⁵ (Fig. 1). These satellite data offer critical advantages in SSA, where only two countries have air-pollution monitoring stations that report to global databases¹², and where chemical transport model-based exposure estimates rely on emission inventories that have high degrees of uncertainty in rural biomass-burning areas^{13,14}.

We model the effects of PM_{2.5} exposure on infant mortality using fixed-effects regression analyses that flexibly account for time-invariant differences in air pollution and mortality across locations, local seasonality in both air quality and mortality, and trending factors or abrupt shocks common to all locations in our sample (see Methods). Because seasonally adjusted variation in PM_{2.5} levels over time at a given location is plausibly exogenous, we propose that this approach isolates the role of poor air quality from other confounding variables that affect mortality risk.

Infant mortality in SSA strongly and linearly increases with PM_{2.5} exposure in our data (Fig. 2a, Extended Data Fig. 3 and Extended Data Table 1). A $10 \mu\text{g m}^{-3}$ increase in PM_{2.5} exposure in the first 12 months of life is associated with a 9.2% increase in infant mortality ($P < 0.01$). We find no qualitative difference between exposure before and after birth at average exposure levels (Fig. 2b), although prenatal associations appear to decline at higher exposure levels. Consistent with recent US evidence¹⁵, we estimate positive associations between PM_{2.5} exposure and mortality at exposure levels below the WHO (World Health Organization)-recommended guideline of $10 \mu\text{g m}^{-3}$ annual average exposure¹⁶ (Extended Data Fig. 4).

Our results are robust to models that use only within-household variation in mortality and PM_{2.5} exposure over time, that allow differential country-level trends in mortality and PM_{2.5} exposure, models that include a large set of additional controls, including temperature, precipitation and household- and child-specific demographic information, and models that use only cross-sectional variation in PM_{2.5} exposure and mortality (Extended Data Figs. 3, 4 and

Africa (higher exposure) versus the rest of Africa (lower exposure). See Extended Data Fig. 2b for countries in each region. **d**, Effect of post-birth exposure on child mortality by terciles of household-level asset wealth, measured as the percentage change in infant mortality per $10 \mu\text{g m}^{-3}$ increase in PM_{2.5} exposure. **e**, Effect of post-birth PM_{2.5} exposure on IMR over time, measured as the percentage change in IMR per $10 \mu\text{g m}^{-3}$ increase in PM_{2.5} exposure.

Extended Data Table 1). Similarly, PM_{2.5} exposure in months 13–24 after birth does not predict mortality in the first year of life (Extended Data Fig. 8i). These findings reduce concerns that results are driven by unobserved factors that are correlated with mortality and PM_{2.5} exposure (for example, spurious time-trending variables), or by the relocation of higher mortality households into locations with poorer air quality (Methods).

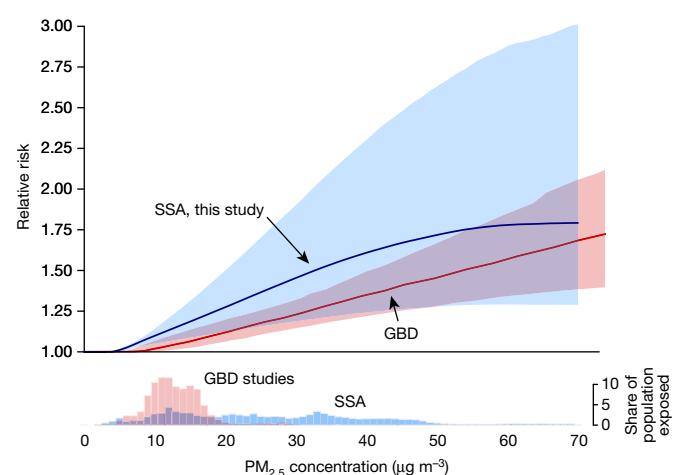


Fig. 3 | Comparing the relative risk curve for all-cause mortality from this study for SSA and the risk curve for respiratory-infection-specific mortality estimated for the Global Burden of Disease (GBD) study. Data for the GBD project were previously published⁴. The GBD acute lower respiratory infection relative risk curve (red) is an integrated exposure response combining point estimates from ambient air pollution studies, indoor air pollution studies and second-hand smoking studies. The relative risk curve estimated in this study (blue) is derived by empirically relating observed births and ambient PM_{2.5} concentrations in SSA (see Methods), with the shaded region representing the bootstrapped 5–95th confidence interval. The histograms show the share of population exposed to different ambient PM_{2.5} concentrations in the regions corresponding to the estimation of each curve. The x axis is restricted to the range of ambient PM_{2.5} concentrations observed in our SSA sample.

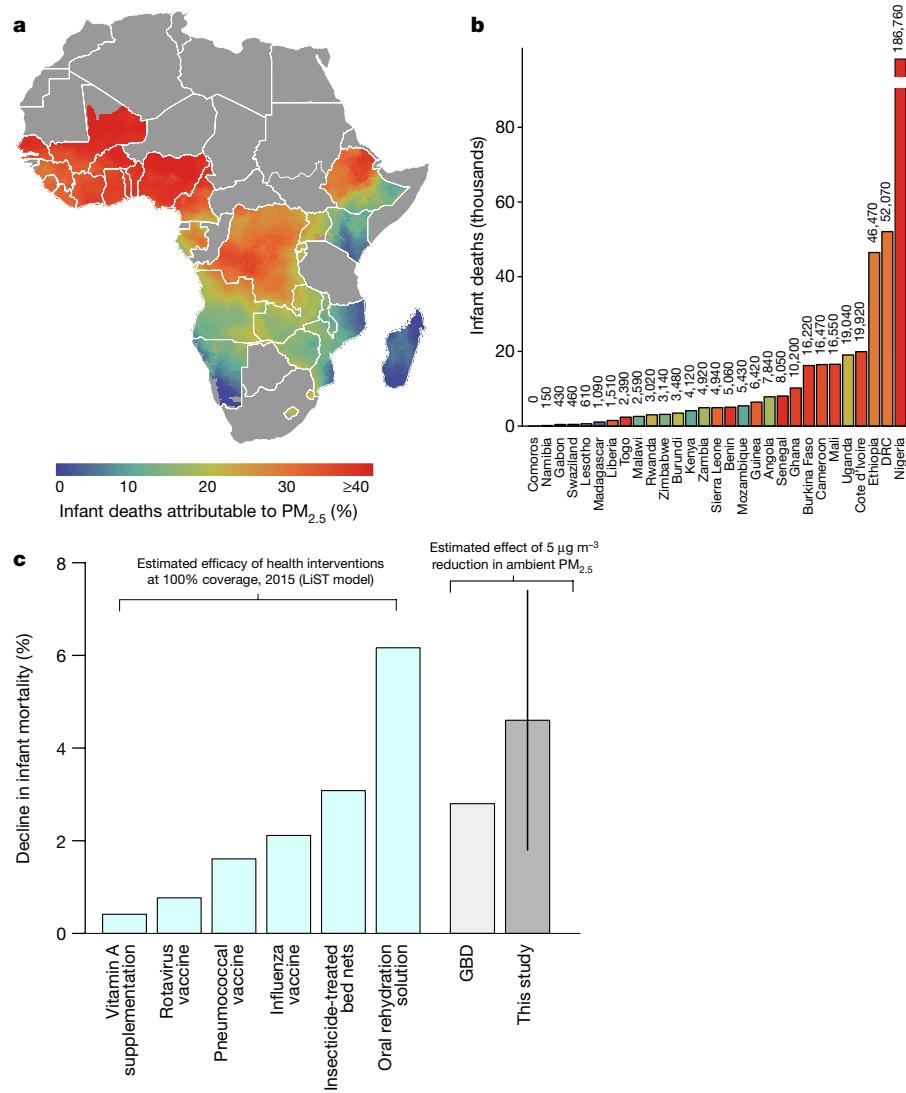


Fig. 4 | Avoided infant deaths from reduced PM_{2.5} exposure.

a, b, Estimated share (a) and number of infant deaths (b) in 30 SSA countries that would have been avoided in 2015 if observed PM_{2.5} levels were reduced to sample minimum exposure of 2 µg m⁻³. The colours in b correspond to the shares shown on the map in a. **c**, Comparison of estimated reductions in infant mortality from achieving 100% coverage

of various health interventions in our study countries from the Lives Saved Tool¹⁹, and estimated reductions in mortality resulting from a PM_{2.5} reduction of 5 µg m⁻³ calculated using the relative risk functions from GBD⁴ or this study. Country outlines were obtained from Global Administrative Areas, version 2.0³⁰. DRC, Democratic Republic of Congo.

Despite differences in the source and level of PM_{2.5} concentrations across African regions (Fig. 1), we find similar associations between PM_{2.5} and mortality when the dataset is restricted to West Africa (where PM_{2.5} levels are higher and partly arise from dust) versus the rest of Africa (where exposures are lower and sources are mainly anthropogenic; Fig. 2c and Extended Data Fig. 2c). Given that much of the PM_{2.5} exposure in West Africa is not a result of local economic activity, these results also suggest that such (unobserved) activity is not biasing our estimated associations between PM_{2.5} exposure and mortality. Similarly, we find no estimated difference in the PM_{2.5}-mortality relationship between households using 'clean' cooking fuels that produce no indoor particulates and households using 'dirty' PM_{2.5}-producing cooking fuels, such as biomass, wood, agricultural residues or dung (Extended Data Fig. 5 and Extended Data Table 2), which suggests that unobserved indoor exposures are not biasing our estimated PM_{2.5}-mortality relationship (see Methods).

A common hypothesis in the environmental health literature is that wealthier households can better avoid the negative health effects of hazardous environmental exposures^{17,18}. However, we do not find evidence that wealth is protective in our setting: associations between PM_{2.5} exposure and mortality risk are similar across wealth terciles

in our data (Fig. 2d), are not moderated by other socio-economic or demographic characteristics (Extended Data Fig. 5 and Extended Data Table 1), and have not declined over time (Fig. 2e and Extended Data Fig. 6).

We use model estimates to construct a relative risk curve for SSA (see Methods). Relative to the mortality risk at the lowest observed exposure levels in our sample (2 µg m⁻³), we estimate a 31% increase in mortality risk at sample median exposure levels (22 µg m⁻³) (Fig. 3). On the basis of this risk curve, we estimate that if PM_{2.5} concentrations in SSA had been reduced to an annual average of 2 µg m⁻³, 22% (95% confidence interval, 9–35%) of infant deaths would have been averted, with the largest reductions in areas with high average exposure (such as most of West Africa; Fig. 4a and Extended Data Fig. 7). We calculate that exposure to PM_{2.5} levels above 2 µg m⁻³ was associated with 449,000 (95% confidence interval, 194,000–709,000) additional infant deaths in 2015 in the 30 study countries alone, with over 40% of these occurring in Nigeria (Fig. 4b).

Although reducing PM_{2.5} concentrations to 2 µg m⁻³ is probably not feasible, substantial reductions in mortality could still be achieved by relatively modest reductions in PM_{2.5} concentrations. We estimate that reducing PM_{2.5} concentrations uniformly by 5 µg m⁻³ at all locations

in our sample countries—a reduction comparable to that achieved by the US Clean Air Act²⁵ (Methods)—would have reduced infant mortality by 4.6% (95% confidence interval, 1.8–7.4%) and avoided 40,000 (95% confidence interval, 20,000–70,000) infant deaths in 2015. This reduction exceeds the estimated mortality reductions that would be obtained if many key child health interventions—including vaccines, nutritional supplementation and insecticide-treated bed nets—were scaled from current levels to 100% population coverage across our study countries¹⁹ (Fig. 4c and Methods). We caution that this comparison does not account for the feasibility or cost effectiveness of achieving these alternative reductions or interventions, but instead provides a key input for future analysis.

Given that current estimates show that around 13% of the mortality of children under 5 years of age in Africa is attributable to lower respiratory infection (LRI)²⁰, our estimate that 22% of infant deaths are attributable to PM_{2.5} exposure would be implausible if LRI were the only channel through which PM_{2.5} exposure affected infant mortality. However, consistent with many recent studies^{21–23} (Methods), we find substantial evidence for effects mediated by non-respiratory channels (Extended Data Fig. 8), including positive associations between in utero PM_{2.5} exposure and neonatal deaths and negative associations with birth weight, and harmful associations between post-birth exposure and both stunting and diarrhoea; all of which are leading causes or risk factors for infant death that overlap only partially with LRI²⁰. We interpret these findings as strong evidence that PM_{2.5} exposure can affect mortality risk through channels other than LRI. As a placebo test, we find no association between PM_{2.5} exposure and child age, sex or likelihood of a multiple birth. Finally, our main estimate of a 9% increase in mortality per 10 $\mu\text{g m}^{-3}$ increase in PM_{2.5} is the same or smaller than the six locality- or country-level quasi-experimental estimates^{24–29} to which our results can be easily compared (Extended Data Fig. 8j).

Our results indicate that risks from PM_{2.5} exposure could be much higher than current global estimates suggest. At median exposure levels in Africa, our estimated relative risk of mortality is double the risk estimated by the GBD at the same exposure level (Fig. 3), and our estimated number of infant deaths associated with ambient PM_{2.5} exposure in our 30 study countries in 2015 is larger than the current GBD estimate of global infant deaths that are attributable to air pollution^{2,6}. Differences could result from our measured associations with all-cause infant mortality, which is broader than LRI-specific mortality used in current global estimates.

Our results also contrast with the common finding that economic development is protective of health^{17,18}, with our data suggesting consistent effects across wealth levels and over time. One potential explanation for this consistency is that we are studying long-term exposure to a pollutant that is small enough to penetrate buildings, making avoidance difficult even for wealthier households.

The greatest impact of poor air quality in our sample is in West Africa, where high PM_{2.5} concentrations include large fractions of dust carried by winds from the Sahara. Although the comparison between West Africa and the rest of the dataset suggests a common exposure-response function, both the particle size distribution below 2.5 μm and the chemical species comprising the PM_{2.5} are unobserved but known to vary widely across sources and regions. More research is needed to characterize these parameters remotely, and to link measurements to prospective epidemiological studies with more detail on both exposures and health outcomes.

Our results indicate that substantial reductions in mortality can be achieved with even modest decreases in ambient PM_{2.5} concentrations. The strong linear relationship between PM_{2.5} and mortality indicates that, even against a high background exposure level, mitigation efforts could deliver large mortality reductions—on par with or exceeding many leading health interventions. This finding is particularly important given the minimal protective benefit of wealth in our data. However, given the varied sources of particulate matter and its precursors across SSA, multi-sectoral and region-specific approaches to reducing exposure burdens may be necessary, and large benefits may

come from developing and adopting protective approaches in dusty regions.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0263-3>.

Received: 14 July 2017; Accepted: 23 May 2018;

Published online 27 June 2018.

1. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525**, 367–371 (2015).
2. Cohen, A. J. et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet* **389**, 1907–1918 (2017).
3. Anenberg, S. C. et al. Impacts and mitigation of excess diesel-related NO_x emissions in 11 major vehicle markets. *Nature* **545**, 467–471 (2017).
4. Burnett, R. T. et al. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. *Environ. Health Perspect.* **122**, 397–403 (2014).
5. van Donkelaar, A. et al. Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environ. Sci. Technol.* **50**, 3762–3772 (2016).
6. Institute for Health Metrics and Evaluation. Global Burden of Disease study 2015 (GBD, 2015) results. <http://ghdx.healthdata.org/gbd-results-tool> (2016). URL.
7. Shindell, D. et al. Simultaneously mitigating near-term climate change and improving human health and food security. *Science* **335**, 183–189 (2012).
8. Zhang, Q. et al. Transboundary health impacts of transported global air pollution and international trade. *Nature* **543**, 705–709 (2017).
9. Ebisu, K., Belanger, K. & Bell, M. L. Association between airborne PM_{2.5} chemical constituents and birth weight—implication of buffer exposure assignment. *Environ. Res. Lett.* **9**, 084007 (2014).
10. West, J. J. et al. What we breathe impacts our health: improving understanding of the link between air pollution and health. *Environ. Sci. Technol.* **50**, 4895–4904 (2016).
11. Burke, M., Heft-Neal, S. & Bendavid, E. Sources of variation in under-5 mortality across sub-Saharan Africa: a spatial analysis. *Lancet Glob. Health* **4**, e936–e945 (2016).
12. Brauer, M. et al. Ambient air pollution exposure estimation for the Global Burden of Disease 2013. *Environ. Sci. Technol.* **50**, 79–88 (2016).
13. Zhang, F. et al. Sensitivity of mesoscale modeling of smoke direct radiative effect to the emission inventory: a case study in northern sub-Saharan African region. *Environ. Res. Lett.* **9**, 075002 (2014).
14. Butt, E. W. et al. The impact of residential combustion emissions on atmospheric aerosol, human health, and climate. *Atmos. Chem. Phys.* **16**, 873–905 (2016).
15. Di, Q. et al. Air pollution and mortality in the medicare population. *N. Engl. J. Med.* **376**, 2513–2522 (2017).
16. Occupational and Environmental Health Team. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2008: summary of risk assessment. *World Health Organization* <http://www.who.int/iris/handle/10665/69477> (WHO, 2006).
17. Patt, A. G. et al. Estimating least-developed countries' vulnerability to climate-related extreme events over the next 50 years. *Proc. Natl. Acad. Sci. USA* **107**, 1333–1337 (2010).
18. Smith, K. R. et al. in *Climate Change 2014: Impacts, Adaptation, and Vulnerability* (eds Field, C. B. et al.) 709–794 (IPCC, Cambridge Univ. Press, 2014).
19. Walker, N., Tam, Y. & Friberg, I. K. Overview of the lives saved tool (list). *BMC Public Health* **13**, S1 (2013).
20. Institute for Health Metrics and Evaluation. GBD Compare data visualization. <http://vizhub.healthdata.org/gbd-compare> (2017).
21. Bell, M. L., Ebisu, K. & Belanger, K. Ambient air pollution and low birth weight in Connecticut and Massachusetts. *Environ. Health Perspect.* **115**, 1118–1124 (2007).
22. Pope, D. P. et al. Risk of low birth weight and stillbirth associated with indoor air pollution from solid fuel use in developing countries. *Epidemiol. Rev.* **32**, 70–81 (2010).
23. Pereira, G., Belanger, K., Ebisu, K. & Bell, M. L. Fine particulate matter and risk of preterm birth in Connecticut in 2000–2006: a longitudinal study. *Am. J. Epidemiol.* **179**, 67–74 (2014).
24. Chay, K. Y. & Greenstone, M. The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. *Q. J. Econ.* **118**, 1121–1167 (2003).
25. Chay, K. Y. & Greenstone, M. *Air Quality, Infant Mortality, and the Clean Air Act of 1970*. Report No. 10053 (National Bureau of Economic Research, 2003).
26. Arceo, E., Hanna, R. & Oliva, P. Does the effect of pollution on infant mortality differ between developing and developed countries? Evidence from Mexico City. *Econ. J.* **126**, 257–280 (2016).
27. He, G., Fan, M. & Zhou, M. The effect of air pollution on mortality in China: Evidence from the 2008 Beijing Olympic games. *J. Environ. Econ. Manage.* **79**, 18–39 (2016).

28. Knittel, C. R., Miller, D. L. & Sanders, N. J. Caution, drivers! Children present: traffic, pollution, and infant health. *Rev. Econ. Stat.* **98**, 350–366 (2016).
29. Cesur, R., Tekin, E. & Ulker, A. Air pollution and infant mortality: evidence from the expansion of natural gas infrastructure. *Econ. J.* **127**, 330–362 (2017).
30. Global Administrative Areas. GADM database of Global Administrative Areas, version 2.0. <https://gadm.org/> (2012).

Acknowledgements We thank D. Lobell, G. McCord, M. P. Burke and W. Schlenker for useful comments and V. Tanutama for research assistance. We thank the Stanford Environmental Ventures Fund and the National Science Foundation (CNH Award #1715557) for funding.

Reviewer information *Nature* thanks R. Black, J. Lelieveld, L. Waller and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.H.N., J.B., E.B. and M.B. designed the research; S.H.N. analysed the data; S.H.N., J.B., E.B. and M.B. interpreted results and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0263-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0263-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Infant mortality data. Data on infant health outcomes are taken from the Demographic and Health Surveys (DHS), nationally representative surveys that are conducted in many low-income and middle-income countries. DHS have a two-stage design, in which a number of clusters are first selected from a list of enumeration areas created in a recent population census, and then households are randomly selected in each of the clusters, and women aged 15–49 years are selected from those households for in-depth surveys. In most survey waves, enumerators use global positioning system devices to collect geospatial information to identify the central point of each cluster's populated area. We used data from all 65 available surveys that were carried out between 2001 and 2015 to reconstruct a village-level time series³¹. Our sample covers 30 countries and includes 990,696 individual birth outcomes (Extended Data Fig. 2). The outcome of interest for this study is infant mortality, which is represented by a dummy variable equal to one when a child was reported to die within the first 12 months after birth. Children who were alive but less than 12 months old at the time of the survey were not included in our sample. The mean infant mortality rate in our sample is 71 deaths per 1,000 births.

Construction of the household wealth measure. The DHS record information on household ownership of a common set of durable assets. In the public distribution files, DHS release a wealth index obtained using a principal components analysis of these household assets and additional services, such as electricity, water supply and floor material, with the index in each survey normalized to that specific survey (that is, wealth quantiles are defined relative to the survey-specific asset distribution). Therefore, although this index enables identification of relative wealth within surveys, it does not allow for comparisons across countries or over time given the within-survey normalization. In order to create a wealth index that could be compared across surveys, we pooled all households with information on the following assets: water source, sanitation facilities, type of flooring, electricity, the number of rooms per person living in the house, and possession of radio, television, phone (landline or cellphone), motorcycle and car. In our dataset, 85% had information on all of these assets. The wealth index was then created using a principal components analysis procedure similar to the survey-specific DHS approach, but normalizing across the entire 65-survey sample rather than within each survey. Further details, including validation and testing of this approach, are available in a previously published paper³².

PM_{2.5} data. We use satellite-derived data on PM_{2.5} compiled by the Atmospheric Composition Analysis Group at Dalhousie University, consisting of annual bias-corrected average surface PM_{2.5} concentrations at 0.01° × 0.01° spatial resolution with global coverage⁵. Building on earlier efforts to predict PM_{2.5} from satellite observations^{33,34}, these data are derived from a suite of satellite-based atmospheric optical depth measurement instruments, including the two MODIS instruments on the Terra and Aqua Satellites, the Multi-Angle Resolution Spectroradiometer (MISR) on Terra and the Sea-viewing Wide Field-of-View Sensor (SeaWiFS) on the SeaStar satellite. These data are combined with aerosol profile measurements from the Cloud–Aerosol Lidar with Orthogonal Polarization (CALIOP) instrument aboard the Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO), and satellite and weather and seasonality data from the GEOS-Chem Chemical transport model to quantify the relationship between column Aerosol Optical Depth (AOD) and surface PM_{2.5} measured at available ground-based stations.

In order to assign PM_{2.5} exposure to individual birth outcomes, we define two exposure periods for each birth: (i) the in utero period encompassing the 9 months before birth and (ii) the post-birth period encompassing the first 12 months of life inclusive of birth month. Given that pollution exposure data are only available annually, we calculate PM_{2.5} exposure in each of these periods as the weighted averages of the annual data, where the weights represent the share of the year that falls into the time period of interest. For example, a child born in the third month of year t would be assigned an in utero pollution exposure of 2/9(exposure in year t) + 7/9(exposure in year $t-1$) and a post-birth exposure of 10/12(exposure in year t) + 2/12(exposure in year $t+1$). The mean PM_{2.5} exposure level for both pre- and post-birth periods is 25.2 $\mu\text{g m}^{-3}$ in our sample.

Ideally, we would account for heterogeneity in the chemical and physical properties of the particulate mix across space and time, but at present, these properties are not directly observed at scales commensurate with overall PM_{2.5}. Moreover, estimates of PM_{2.5} properties rely heavily on aerosol models and emissions inventories³⁵, which are known to be highly uncertain in biomass-burning regions such as SSA^{13,14}. As an example, the Atmospheric Composition Analysis Group offers a version of the PM_{2.5} dataset used in this analysis with dust and sea salt aerosols removed. However, this 'dust-free' version is not based on observational partitioning methods, but simply scaled based on an emissions inventory. As such, we use only the observationally constrained full PM_{2.5} dataset for our analysis, but note that finer-grained observations of the chemical and physical properties of aerosol

particulate matter could be used in the future to understand whether and to what extent impacts change with PM_{2.5} chemical composition and size distribution.

Empirical approach. We model the relationship between infant deaths y and PM_{2.5} exposure using a least squares linear probability model:

$$y_{icnmt} = f(\text{PM}_{icnt}^b, \text{PM}_{icnt}^a) + \lambda \mathbf{X}_{icnmt} + \mu_c + \delta_t + \eta_{nm} + \varepsilon_{icnmt} \quad (1)$$

where i indexes child, c indexes survey cluster (that is, village), t indexes birth year, and nm indexes country and month. PM_{icnt}^b and PM_{icnt}^a refer, respectively, to PM_{2.5} exposure in the 9 months before and 12 months after birth. \mathbf{X}_{icnmt} is a vector of additional controls potentially relevant to the relationship between PM_{2.5} and infant mortality, including household and individual characteristics, such as child sex, birth order, age of the mother, education of the mother, type of cooking fuel used at home and our asset-based wealth index, as well as time-varying climate variables such as temperature and precipitation. We do not include wealth as a control in our main results because the information is not available for the full sample, but show that results are unchanged upon its inclusion (Extended Data Table 1b). μ_c , δ_t and η_{nm} are DHS cluster, birth year and country-month effects, respectively. The cluster effects control for time-invariant cross-village differences (for example, higher or lower average mortality levels), year effects control flexibly for trends or abrupt shocks common to all locations (for example, macroeconomic shocks or declines in mortality over time), and country-month effects control for seasonality in infant mortality and PM_{2.5} exposure. In order to make nationally representative survey data representative of the entire 30-country sample, we follow a previous publication³⁶ and weight observations by the product of country-specific household survey weights (supplied by DHS) and country population weights; however, our results are insensitive to dropping the weights.

GBD estimates suggest that $f(\cdot)$ is nonlinear, with marginal effects of PM_{2.5} exposure declining at higher exposure levels⁴. To explore potential nonlinear responses to PM_{2.5} in our data, we estimate flexible versions of $f(\cdot)$, including higher-order polynomials and restricted cubic splines. However, we find that flexible models for post-birth exposure provide roughly the same shaped response function as simple linear models (Extended Data Fig. 3), and that higher-order polynomial terms for the post-birth period are not statistically significant in the full sample (Extended Data Table 1). We therefore model post-birth PM_{2.5} linearly in our main specification and in the calculation of attributable deaths; including higher-order terms generally steepens the relationship (Extended Data Fig. 3) and yields higher attributable death estimates. The quadratic term for the 9-month in utero period is statistically significant, however, and thus our main specification adopts $f(\cdot)$ quadratic in in utero PM_{2.5} exposure (PM^b) and linear in post-birth PM_{2.5} exposure (PM^a):

$$f(\text{PM}_{icnt}^b, \text{PM}_{icnt}^a) = \beta_1 \text{PM}_{icnt}^b + \beta_2 (\text{PM}_{icnt}^b)^2 + \beta_3 \text{PM}_{icnt}^a \quad (2)$$

To study whether the impacts of PM_{2.5} change over time or by wealth level, we interact linearized post-birth exposure with dummy variables for wealth quantile or for year of survey:

$$y_{icnmt} = \sum_d \beta_d (I_d \text{PM}_{icnt}^a) + \mu_c + \delta_t + \eta_{nm} + \varepsilon_{icnmt} \quad (3)$$

where I_d is a dummy variable for whether observation i falls into bin d . The β_d coefficients provide the marginal effect of a $1 \mu\text{g m}^{-3}$ separately for each bin (wealth quantile or time period). For the wealth estimates, we focus on terciles of the wealth index.

The goal of the fixed effects in equations (1) and (3) is to isolate variation in PM_{2.5} exposure from other time-invariant, seasonally varying or time-trending factors that could be correlated with mortality. For instance, by including cluster-fixed effects and thus using only within-village deviations in PM_{2.5} and mortality over time, our approach accounts for time invariant unobservables that could be correlated with both PM_{2.5} exposure and mortality risk at the cluster level (for example, if villages with lower PM_{2.5} exposure also happen to be wealthier). Because we observe more than one birth for most mothers, our data allow an even more stringent test on the potential role of time-invariant household unobservables. In particular, we can include a mother fixed effect in equation (1) (the cluster fixed effects thus drop out), and in this design, the effects of PM_{2.5} exposure on mortality derive from comparing whether a child born to a given mother during a period of high PM_{2.5} exposure is more or less likely to survive relative to a child born to that same mother during a period of lower PM_{2.5} exposure. This within-mother variation eliminates the common concern in pollution exposure studies that households with different levels of pollution exposure could be inherently different in unobservable ways (for instance if wealthier, lower-mortality mothers choose to live in areas of lower pollution exposure). Our results using mother fixed effects are very similar to results using cluster fixed effects (Extended Data Fig. 3), again providing strong evidence that our results are not being driven by time-invariant

unobservables. Whether or not we control directly for a broader set of individual and household characteristics X_{icmt} also does not change our results (Extended Data Table 1).

Similarly, overall trends in mortality and $PM_{2.5}$ exposure over time are taken out by the year fixed effects (and in robustness checks, by time trends or country-by-year fixed effects, see Extended Data Fig. 3), helping to reduce concerns that the effects of $PM_{2.5}$ that we estimate are driven by common time-trending unobservables. However, common time effects would not account for local time-varying factors that could be correlated with both $PM_{2.5}$ and mortality. Of particular importance is rainfall: rainfall reduces $PM_{2.5}$ in the atmosphere, and rainfall could be plausibly negatively or positively correlated with mortality—positively if higher rainfall led to favourable conditions for transmission of vector-borne disease (for example, malaria) or negatively if higher rainfall led to greater local food availability and thus lower mortality. To account for this possibility, we control directly for meteorological conditions using high-resolution remote-sensing based gridded data on precipitation and temperature^{37,38}, and find that our results are unaffected (Extended Data Table 1).

A related concern is that local economic activity could be associated with both local $PM_{2.5}$ levels and mortality. We believe that this is less important in our setting for two reasons. First, for large parts of our sample (particularly in West Africa), much of the variation in $PM_{2.5}$ is driven by wind-borne dust, which is unrelated to local economic activity. Second, because economic growth is most likely associated with both higher $PM_{2.5}$ levels (from industrial or agricultural activities) and lower infant mortality³⁹, then this would bias our results towards zero.

A final concern is that although equation (1) identifies impacts using variation over time in $PM_{2.5}$ exposure that is plausibly orthogonal to other determinants of mortality, individuals over the long run might adapt to differing levels of average pollution exposure in a way that is not picked up in a time series—for example, they might undertake defensive investments or learn how to limit exposures or reduce their consequence. Panel models that use inter-annual $PM_{2.5}$ variation might then overstate the harm of $PM_{2.5}$ exposure, because variation around local averages is harder to anticipate and adapt to. Although cross-sectional models that relate location-average mortality to location-average pollution exposure are subject to bias concerns from omitted variables and are considered unreliable for estimating causal effects, they arguably have the benefit of accounting for general forms of longer run adaption. We find that panel and cross-sectional models indicate surprisingly similar responses of mortality to pollution (Extended Data Fig. 4a), suggesting limited adaptation over the longer run.

Calculating relative risk and excess deaths attributable to pollution. Using the full $y(\cdot)$ function estimated in equation (1), we calculate the relative risk (RR) at a given $PM_{2.5}$ exposure level z as the predicted values from the full model evaluated at $PM_{2.5} = z$, divided by the predicted values from the full model evaluated at $PM_{2.5} = 2$:

$$RR(z) = \frac{y(z)}{y(2)} \quad (4)$$

where $2 \mu\text{g m}^{-3}$ represents the minimum exposure level observed in our data. Our approach to defining the lower bound for risk is thus similar to the approach in the GBD⁴, who define the lower bound in their relative risk curve as the minimum $PM_{2.5}$ exposure level observed in a constituent cohort study ($z = 5.8 \mu\text{g m}^{-3}$). Although estimates are imprecise at very low exposure levels due to limited sample sizes, both flexible splines and piecewise linear functions suggest that, in our data, mortality is increasing with $PM_{2.5}$ even at the lowest observed exposure levels in our data, and for this reason we set our ‘reference’ risk level to $z = 2 \mu\text{g m}^{-3}$.

To calculate relative risk across the entire range of observed $PM_{2.5}$ levels in our data and for all geographical locations, relative risk is calculated for every birth observation at its observed post-birth $PM_{2.5}$ concentration and then averaged to the cluster level. To calculate the relative risk curve in Fig. 3, we divide the data into $5 \mu\text{g m}^{-3}$ $PM_{2.5}$ bins, calculate the average relative risk within each bin, and then fit a flexible locally weighted polynomial to these estimates. Confidence intervals are obtained by bootstrapping equation (1) 1,000 times, sampling clusters with replacement, and recalculating equation (4) for each bootstrapped $y(\cdot)$. The 5–95th confidence interval is then the 5th and 95th percentiles of these 1,000 estimates at each point in the $PM_{2.5}$ distribution. Measurement error in our outdoor $PM_{2.5}$ measures, which is estimated to be roughly classical⁵, will lead to attenuation bias in our coefficient estimates in equation (1), and thus mean that our relative risk curve is biased towards zero. We discuss an alternate source of non-classical measurement error—the error related to unobserved indoor air pollution exposure—in ‘Indoor versus outdoor $PM_{2.5}$ ’.

We calculate the share of infant deaths attributable to $PM_{2.5}$ exposure in each DHS location i as:

$$S_i = 1 - \frac{y(2)}{y(z_i)} = 1 - \frac{1}{RR_i} \quad (5)$$

The average share of $PM_{2.5}$ -attributable deaths across the sample is then calculated as the population-weighted average across DHS locations, using high-resolution gridded data on birth counts from WorldPop⁴⁰ as weights. For each country, WorldPop produces a $100 \times 100 \text{ m}^2$ grid of birth counts that, when aggregated, are consistent with UN estimated country totals.

To generate the country-wide surfaces shown in Fig. 4, we apply the relative risk curve in Fig. 3 to all locations in our sample countries, using grid-level observed $PM_{2.5}$ levels in 2015. Mean exposure levels in 2015 were $30 \mu\text{g m}^{-3}$, or about $5 \mu\text{g m}^{-3}$ higher than overall sample average exposure, and thus the share of attributable deaths in 2015 shown in Fig. 4 is a little above the 22% average that we calculate for the full sample.

Finally, to calculate the total additional infant deaths attributable to $PM_{2.5}$ in 2015, we calculate for each location i :

$$ED_i = B_i IMR_i S_i \quad (6)$$

where B_i is the estimated number of births in location i in 2015 from WorldPop, IMR_i is the estimated average infant mortality rate (IMR) between 2005 and 2015 as calculated by applying previously published methods¹¹ to the more recent infant mortality data used in this study (map shown in Fig. 1c), and S_i is the share of mortality attributable to $PM_{2.5}$ as calculated above. The total attributable deaths across our sample countries in 2015 is then the sum of ED_i over all locations. Confidence intervals are calculated as above by recalculating ED_i across bootstrapped estimates of S_i .

Comparison to GBD. In a recent study⁴, a global integrated response function was derived that relates ambient $PM_{2.5}$ exposure to the relative risk of acute lower respiratory infection in infants (reproduced in Extended Data Fig. 1a). (A recent update² to our knowledge did not provide age-specific response functions to which our estimates can be compared.) To develop the global relative risk estimates, the GBD authors relied on the available literature (see the previous study⁴ for datasets and references) at the time, which consisted of: (i) 4 studies that measured the effect of ambient exposures on health outcomes, all from developed countries and with average ambient exposures below our African sample median (Extended Data Fig. 1a); (ii) 23 studies that measured the effect of second-hand smoking on health outcomes, all of which were assigned the same ‘ambient’ exposure level of $50 \mu\text{g m}^{-3}$, because true exposures were unobserved; and (iii) 1 study of household carbon monoxide exposure on child respiratory outcomes in Guatemala, for which $PM_{2.5}$ exposures had to be inferred for a large proportion of the sample, and for which counterfactual (minimum) ambient exposures were substantially higher than in all ambient studies. As shown in Extended Data Fig. 1b, of these 28 studies, 8 were in developing countries, and only one in Africa, and the median study sample was $n = 1,250$ individuals.

In comparison, our study (i) observes nearly one million individuals, more than the combined sample size of the 28 studies of acute lower respiratory infection described in the previous study⁴; (ii) directly studies the effect of ambient exposure on health outcomes in a developing country setting using quasi-experimental variation in $PM_{2.5}$ exposure to estimate health effects; (iii) uses a single empirical approach and data source to estimate a relative risk function across a broad range of $PM_{2.5}$ exposures, meaning that differences in measured responses across exposure levels cannot be attributed to differences in empirical approach or study design in different locations; and (iv) in keeping with growing literature suggesting non-respiratory effects, does not assume that the only pathway linking $PM_{2.5}$ to overall health outcomes is respiratory infection.

Differences in estimated relative risk functions between our study and Burnett et al. are shown in Fig. 3 and discussed in the main text. Given the methodological and locational differences between our study and the previous study⁴, we emphasize that differences in relative risk between the two studies at specific exposure levels cannot be interpreted as providing evidence on (for example) the relative damages caused by ambient $PM_{2.5}$ exposure versus exposure to second-hand smoke or indoor air pollution.

To compare our estimates of attributable deaths to those of the GBD, we recalculate equation (5) using the relative risk function published previously⁴, using the same grid-level $PM_{2.5}$ and population numbers that we used to generate our attribution estimates, but keeping the previously published⁴ counterfactual exposure of $5.8 \mu\text{g m}^{-3}$. Using the previously published relative risk function⁴ and this higher counterfactual exposure, we estimate that 13% of infant deaths in our sample are attributable to $PM_{2.5}$ exposure. This estimate is contained within the confidence interval for our main estimate of 22% (9–35%).

An important difference between our attribution calculations and previous study⁴ is that the latter uses a counterfactual $PM_{2.5}$ exposure (that is, theoretical minimum risk exposure level) of $5.8\text{--}8.0 \mu\text{g m}^{-3}$, which were the lowest and fifth percentiles of exposure in their reference studies (although we

note that a recent update by the GBD team² now uses $2.4 \mu\text{g m}^{-3}$ as the counterfactual exposure). As noted in the previous study⁴, these thresholds were chosen based on “(a) the availability of convincing evidence from epidemiologic studies that support a continuous reduction in risk of disease to the chosen distribution, and (b) a distribution that is theoretically possible at the population level.” We follow this logic and set our lower threshold at $2 \mu\text{g m}^{-3}$, as this is the minimum exposure level measured in our sample, and because we find evidence of linear effects of $\text{PM}_{2.5}$ on mortality at levels below $10 \mu\text{g m}^{-3}$ (Extended Data Fig. 4a). However, our choice of a lower counterfactual exposure, while arguably appropriate in our setting, could generate some of the difference that is observed between our attribution estimates and GBD’s. To understand the importance of this choice, we re-calculate our population-weighted attributable deaths under different counterfactual exposures, from $2 \mu\text{g m}^{-3}$ up to $10 \mu\text{g m}^{-3}$ (Extended Data Fig. 7). We find that the share of infant deaths attributable to $\text{PM}_{2.5}$ exposure ranges from 22% with our counterfactual of $2 \mu\text{g m}^{-3}$ to 13% with a counterfactual of $10 \mu\text{g m}^{-3}$. Using previously published⁴ minimum counterfactual of $5.8 \mu\text{g m}^{-3}$, our estimate of attributable deaths becomes 18%, compared to the 13% that we would calculate using the relative risk function of the previous study⁴ (as described above).

In order to calculate a comparable estimate of additional deaths attributable to air pollution in 2015 from the GBD studies, data were downloaded from the GBD results tool⁶ for 2015 and infant deaths attributable to air pollution were summed over the 30 countries in our sample. The ‘air pollution’ category in GBD includes the categories ‘ambient particulate matter pollution’, ‘ambient ozone pollution’, and ‘household air pollution from solid fuels’. We compare our estimates of $\text{PM}_{2.5}$ -attributable death to GBD estimates of the total overall ‘air pollution’-attributable death estimates, which will be an upper bound on GBD-attributed deaths from $\text{PM}_{2.5}$ exposures specifically, and thus provide the most conservative possible comparison. This comparison is also preferable on physical grounds: it is both statistically and functionally difficult to distinguish indoor and outdoor air pollution exposures in rural biomass-burning regions, especially since most of the $\text{PM}_{2.5}$ from household cooking and fires makes its way outdoors⁴¹. The inclusion of ‘ambient ozone pollution’ is a small effect because ozone-related mortality⁴² is typically an order of magnitude lower than for $\text{PM}_{2.5}$, and it is also difficult to distinguish from $\text{PM}_{2.5}$ -related impacts because both are often present in local pollution.

The GBD results tool indicates air pollution-attributable neonatal deaths of 126,000 in our 30 countries in 2015 (range = 73,000–198,000), 150,000 in all of SSA (118,000–184,000) and 294,000 globally (234,000–350,000). Our estimate of 449,000 attributable deaths in 2015 is thus $3.6 \times$ higher than the GBD estimate for the same countries. Revising the attributable death estimates upward in our 30 countries would result in an additional attributable 323,000 deaths, which would represent a more than doubling of the global estimated attributable deaths to air pollution.

Alternatively, we can apply the approach described in equation (6) to the integrated response curve developed previously⁴. This approach produces an estimate of 336,000 attributable deaths in 2015, closer to—but still substantially smaller than—our estimate of 449,000 attributable deaths. Our finding that lower ranges of $\text{PM}_{2.5}$ exposure are more harmful to infant health than previously thought is one of the factors that contributes to the difference in estimates. For example, only 11% of attributable deaths (38,000) estimated using previously published response curve⁴ occurred in locations with lower than median ($27 \mu\text{g m}^{-3}$) $\text{PM}_{2.5}$ exposures whereas 18% (80,000) of attributable infant deaths estimated using our methods occurred in these relatively lower $\text{PM}_{2.5}$ exposure areas.

Comparison to other health interventions. We compared the estimated effectiveness of a given reduction in $\text{PM}_{2.5}$ exposure from both our model and the previous study⁴, to the estimated effectiveness of other important health interventions based on estimates from the Lives Saved Tool¹⁹ (LiST; <http://www.livessavedtool.org/>, accessed 20 September 2017). LiST is a model designed to estimate the effect of scaling up health and nutritional interventions on child and maternal health. For each intervention of interest, LiST takes as input country-specific demographic information, cause-of-death data, current intervention coverage, and data from randomized controlled trials and quasi-experiments on intervention efficacy. It then combines this information to estimate the effectiveness (in terms of reduced mortality) of scaling the intervention to a desired level of population coverage.

We used LiST to estimate the mortality impacts of scaling the following interventions from baseline 2015 coverage rates to full (100%) coverage: vitamin A supplementation, selected vaccines (rotavirus, pneumococcal, influenza), insecticide-treated bed nets and oral rehydration solution. We did this separately for each of the 30 countries in our sample (except Tanzania, which was not included in the LiST database), using the default demographic datasets provided in LiST. Baseline population-weighted coverage in 2015 for the interventions were: vitamin A supplementation (73%), rotavirus vaccine (32%), pneumococcal vaccine (54%), influenza B vaccine (74%), insecticide-treated bed nets (54%) and oral rehydration

solution (37%). The LiST-estimated baseline infant mortality across these countries in 2015 was 56.7 deaths per 1,000 live births. Country-specific estimates of mortality reductions due to scaling each intervention to 100% population coverage were then averaged (weighting by population) to produce the overall estimates reported in Fig. 4.

We compared the LiST estimates to the estimated effect of a $5 \mu\text{g m}^{-3}$ $\text{PM}_{2.5}$ reduction, using both our model (equation (5)) and the previously published relative risk function⁴ using the approach described above. A $5 \mu\text{g m}^{-3}$ reduction is roughly equivalent to the estimated reduction in $\text{PM}_{2.5}$ induced by the 1970 Clean Air Act (CAA) in nonattainment counties in the United States; although $\text{PM}_{2.5}$ was not routinely measured in the US until the 1990s, the CAA is estimated⁴³ to have reduced total suspended particulates (TSP) by 20–25 $\mu\text{g m}^{-3}$, and evidence from multiple sites in North America^{44,45} suggest roughly 25% of TSP by mass is $\text{PM}_{2.5}$, meaning the CAA led to $\text{PM}_{2.5}$ reductions on the order of $5 \mu\text{g m}^{-3}$.

We emphasize that our comparison of the effectiveness of $\text{PM}_{2.5}$ reductions to that of other health interventions abstracts from the policy, technical and/or financial realities of implementing these reductions or interventions. As with the LiST model, our purpose is rather to provide a basis for further exploration of the comparative feasibility and cost effectiveness of alternate interventions.

Indoor versus outdoor $\text{PM}_{2.5}$. One concern with our results is that although we purport to measure the relationship between outdoor air pollution and infant health, infants could also be exposed to indoor air pollution, and this unobserved exposure could bias our estimates. Here we quantify the likely sign and magnitude of the bias, and show that unobserved indoor air pollution probably leads us to underestimate the effect of outdoor air pollution.

Indoor and outdoor air pollution have traditionally been treated as distinct public health threats, with the separation largely reflecting the difference in pollution sources (as opposed to biological impact mechanisms) and the anticipated differential impacts of technologies or policy responses aimed at those sources. The main ‘indoor’ source of aerosol particulate matter in SSA is cooking using solid, unprocessed fuels such as wood, dung, agricultural residues, charcoal, and coal. Nearly three billion people worldwide still depend primarily on such fuels (rates are extremely high—85%—in our sample), and women and young children tend to be disproportionately exposed to cooking-related emissions given the gendered breakdown of domestic tasks in much of the world. This stands in contrast to outdoor $\text{PM}_{2.5}$ sources such as electric power generation, transportation, and open biomass burning, which are assumed to affect nearby populations more homogeneously.

But while these indoor and outdoor emissions sources may be distinct, the separation of exposures to those emissions is difficult. The basic connections between indoor and outdoor environments are well established: (i) much of the $\text{PM}_{2.5}$ that originates indoors is transported outdoors through chimneys, windows, and doors, meaning that in rural areas in developing regions, cooking-related emissions can actually drive outdoor concentrations⁴¹; and (ii) absent sophisticated filtering, outdoor concentrations represent the lower limit for indoor exposures, because air must be exchanged periodically. Although very few studies in SSA feature simultaneous indoor and outdoor measurements (and most measure PM_{10} and/or carbon monoxide (CO) instead of $\text{PM}_{2.5}$), they support several generalized findings: that a significant amount of cooking happens outdoors or in cooking areas separated from the rest of the house, that outdoor pollutant concentrations track indoor (cooking-related) concentrations⁴⁶, that concentrations in the immediate cooking area/cookstove plume spike much higher, with concentrations rapidly falling off with distance, and that areas elsewhere in the house can be relatively protected and personal exposures can vary widely^{47–49}.

We highlight the literature from SSA, albeit small, as much of the indoor/outdoor literature that includes cooking emissions has focused on highland areas in China, India and Central and South America, where indoor heating is a key service provided by indoor combustion and so ventilation conditions can be very different^{50,51}. These studies also highlight that concentrations within houses and nearby areas vary markedly during cooking hours and across seasons⁵². The indoor/outdoor literature from developed countries focuses on how well buildings (which must nevertheless exchange air with the outdoor environment) keep out pollutants, including $\text{PM}_{2.5}$. These studies⁵³ highlight that indoor:outdoor concentration ratios span 1 when windows are open and there is direct air exchange, as is the case in most of SSA/our DHS sample.

A perfect exposure metric would integrate indoor and outdoor exposures over time spent in the two environments (or weight an average of the two by relative time spent in each location). We do not have indoor concentration data for our study regions or individual exposure data; as noted above, very few simultaneous indoor and outdoor measurements exist anywhere, and especially in SSA. We therefore proxy for integrated $\text{PM}_{2.5}$ exposure by average ambient (outdoor) concentrations, derived from satellites, ground monitors, and chemical transport models, as described in ‘ $\text{PM}_{2.5}$ data’. Therefore although infant mortality rates

undoubtedly are a function of total integrated $PM_{2.5}$ exposure, we estimate the response only to the observable outdoor portion. Here we assess the extent to which this approximation is valid.

We can write total exposure (PM_{tot}) as a weighted average of indoor and outdoor exposures, where a_{out} and a_{in} are the fraction of time exposed to outdoor and indoor concentrations, respectively:

$$PM_{tot} = a_{out}PM_{out} + a_{in}PM_{in} \quad (7)$$

We expect that IMR is a function of total exposure (that is, indoor plus outdoor), but that the health impact of a given amount of either indoor or outdoor exposure is the same. Thus the correct exposure model, assuming linearity for instructive purposes, would be:

$$IMR = IMR_0 + \beta(a_{out}PM_{out} + a_{in}PM_{in}) \quad (8)$$

with β being the 'true' response of IMR to $PM_{2.5}$ exposure. Noting that $a_{out} + a_{in} = 1$, we can write:

$$IMR = IMR_0 + \beta(PM_{out} + a_{in}(PM_{in} - PM_{out})) \quad (9)$$

If we assume a general relationship between PM_{in} and PM_{out}

$$PM_{in} = \delta_0 + \delta PM_{out} \quad (10)$$

we can rewrite equation (9) as:

$$IMR = IMR_0 + \beta(PM_{out} + a_{in}(\delta_0 + \delta PM_{out} - PM_{out})) \quad (11)$$

which simplifies to:

$$IMR = (IMR_0 + \beta a_{in} \delta_0) + \beta(1 + a_{in}(\delta - 1))PM_{out} \quad (12)$$

We only observe PM_{out} and thus estimate the coefficient on the model:

$$IMR = IMR_0 + \alpha(PM_{out}) \quad (13)$$

From this regression (which is analogous to our main regression in equation (1)), we recover estimates of

$$\hat{\alpha} = \beta(1 + a_{in}(\delta - 1))$$

The key question is the extent to which $\hat{\alpha}$ diverges from the true response β . The fact that we do not observe PM_{in} , δ_0 , δ , a_{out} , or a_{in} leads to several possibilities for α . Our estimate $\hat{\alpha}$ will clearly be unbiased when $a_{in} \approx 0$, the case if the child is not exposed to indoor concentrations, either because they are not indoors, or (most practically) because cooking happens in a different location (for example, a separate kitchen) and they are effectively protected from those emissions while indoors. Although there is evidence that young children often have less exposure to indoor air pollution than others in the family (particularly adult females)⁴⁷, and some countries represented in our sample have fairly high rates of outdoor cooking, we nevertheless view it as unlikely that most children in our sample are unexposed to indoor pollution.

If $a_{in} > 0$, then the extent of bias depends on δ . Our estimates $\hat{\alpha}$ recover the true effect β when $\delta = 1$, that is, when PM_{in} scales exactly with PM_{out} (up to a constant offset δ_0). When $0 < \delta < 1$, our estimates of α underestimate the true effect of $PM_{2.5}$ on health; the opposite is true if $\delta > 1$. There are two ways to think about parameter δ . The first is at the household level. In a household model, δ_0 can be thought of as the time-averaged concentration from indoor emission sources, and δ can be thought of as the time-averaged steady-state balance of total $PM_{2.5}$ mass transport from outdoor-to-indoor compared to $PM_{2.5}$ mass transport from indoor-to-outdoor. Because some air must be exchanged between indoor and outdoor environments, $\delta > 0$, and although the volume of air exchanged will be equal, the total mass of $PM_{2.5}$ transported in either direction can differ. A scenario in which $\delta > 1$ means that $PM_{2.5}$ is trapped and builds up indoors; this leads to estimates of α that exceed the true β . This scenario is highly unlikely in an environment such as SSA with unsealed buildings. Instead, the more likely scenario is that $\delta \leq 1$. $\delta = 1$ would be the case in households that do not have indoor $PM_{2.5}$ sources (for example, clean cooking fuels), so indoor concentrations = outdoor concentrations, and air volumes are cycled back and forth. $\delta < 1$ for households with indoor emissions sources that are ventilated to the outdoors: the total mass transported outdoors is larger than the mass transported indoors (a higher concentration of $PM_{2.5}$ in the ventilated air than in outdoor air exchanged for it). The household-level interpretations of δ and δ_0 are summarized in Extended Data Table 2a.

For households using clean cooking fuels, there is no indoor $PM_{2.5}$ source ($\delta_0 = 0$), so the possible scenarios are the top row of Extended Data Table 2a. For houses using dirty cooking fuels ($\delta_0 > 0$, that is, there is an indoor source contributing to a steady-state indoor concentration that is unrelated to outdoor levels), the possible scenarios are the bottom row of the table. As described above, so long

as $\delta > 0$, the bias in $\hat{\alpha}$ does not depend on whether or not the household has an indoor source of $PM_{2.5}$ (for example, it uses only clean cook fuels). Cells 3 and 6 ($\delta > 1$) of Extended Data Table 2a are highly unlikely in steady-state as they would imply that $PM_{2.5}$ mass from outdoors is being transported indoors and concentrating there. Cell 1 is physically impossible without an indoor source of $PM_{2.5}$, and cell 5 would be an idiosyncratic case in which ventilated indoor $PM_{2.5}$ concentrations were exactly matched by outdoor concentrations (this might be approximately the case for cooking emissions ventilated immediately via a chimney just outside the house, where they are pulled back in again). So from a household model, we would expect $\hat{\alpha}$ to be unbiased for households with clean cooking fuels (no indoor $PM_{2.5}$ generation), and underestimated for households with indoor $PM_{2.5}$ sources.

We can evaluate this prediction in our data, given that DHS data do provide information on the use of cook fuels of households for a subset of households. We define 'clean' cook fuel households as those who cook with natural gas, biogas, liquefied petroleum gas (LPG) or electricity, and 'dirty' fuel households as those cooking with anything else. Consistent with the prediction above, point estimates suggest larger effects of $PM_{2.5}$ exposure on clean fuel households compared to dirty fuel households, although we cannot reject that the estimates are the same given the wide confidence intervals (Extended Data Fig. 5).

The second way to think about δ is at the aggregate level, or the population relationship between indoor and outdoor $PM_{2.5}$ concentrations. Intuitively, we would expect a positive correlation between indoor and outdoor $PM_{2.5}$ concentrations in aggregate, because outdoor $PM_{2.5}$ penetrates porous buildings and households ventilate indoor emissions. Bias in $\hat{\alpha}$ would only occur if that relationship changed across levels of outdoor $PM_{2.5}$ exposures. This would imply that, for example, at higher outdoor $PM_{2.5}$ concentrations, indoor $PM_{2.5}$ is less well-ventilated, or that at higher outdoor concentrations, homes are trapping and concentrating indoor $PM_{2.5}$.

To explore these aggregate-level relationships in our data, we aggregate DHS and ambient pollution data to the cluster level, restrict the sample to within the survey year (given that survey questions ask about current fuel use) and test the relationship between the percentage of households in a cluster using 'clean' fuels and ambient $PM_{2.5}$ levels. Results are shown in Extended Data Table 2b. We find that, on average, clusters using entirely clean cooking fuels have lower ambient $PM_{2.5}$ levels, both across the full sample and (more importantly) when restricted to only clusters with at least some clean cook fuel use. To verify that this isn't simply result of unobserved variables—for example, locations with access to clean cooking fuels might also have reduced open biomass burning, or better (cleaner) electric power generation, or lower transportation-related $PM_{2.5}$ emissions—we conduct separate regressions on urban and rural clusters. We find no difference in the impact of cluster-level clean cooking fuel penetration on ambient levels in rural versus urban DHS clusters (where we would expect non-cooking emissions profiles to differ). This provides additional evidence that our ambient average metric proxies well for overall exposure, and that the relationship is not driven by non-cooking emissions, because cooking-related emissions that originate indoors are ultimately reflected in outdoor average concentrations.

These three pieces of evidence—that the basic physics of air flow suggest that $0 < \delta \leq 1$, that 'clean' fuel households have higher point estimates of $PM_{2.5}$ effects, and that outdoor concentrations appear to reflect indoor exposures—suggest that, if anything, unobserved indoor $PM_{2.5}$ exposures likely bias our main estimates down. We thus interpret our main estimates as conservative.

Nevertheless, infants could be extremely vulnerable to quick but marked spikes in indoor particulates during, for example, stove lighting. The existing literature for SSA highlights the high temporal variability in pollutant emissions from cooking, and the tremendous spatial heterogeneity in concentrations over short spatial scales⁴⁷. Simultaneous measurement of direct exposure for individuals, in addition to indoor and outdoor concentrations, is a key area for future research that would help to clarify the role of higher frequency spatiotemporal variation in concentrations/exposures, and allow for comparison of integrated (versus average) exposure.

Impact channels and effect size plausibility. One of our main results is that 22% of infant deaths in our sample can be attributed to $PM_{2.5}$ exposure above $2 \mu\text{g m}^{-3}$. This estimate is larger than current GBD estimates of the total child mortality burden of LRI in SSA; the online GBD tool²⁰ estimates that 12.8% of infant deaths in SSA in 2015 were due to LRI. The difference between estimates makes our results seem implausibly large if (i) the GBD estimates are correct and (ii) LRI are the only cause of death linking $PM_{2.5}$ exposure and infant mortality.

As a first point, we note that if the GBD counterfactual $PM_{2.5}$ concentration of $5.8 \mu\text{g m}^{-3}$ is used to define our 'clear air' baseline, then our estimate of attributable deaths would be reduced to 18%. So different clear-air counterfactuals could result in some of the difference between our estimate and GBDs.

Second, when we use the exposure-response function that was the basis of the 2015 GBD estimates⁴, we calculate for our SSA sample that 13% of infant deaths are attributable to $PM_{2.5}$ exposure (see above). This estimate is within our estimated 95% confidence interval and suggests that even in the GBD, LRI alone cannot

account for the mortality attributable to $PM_{2.5}$, unless we are willing to attribute all deaths from LRI to $PM_{2.5}$, an unlikely scenario given the multiple agents that cause LRI in young children. In other words, even the GBD data appear to indicate that $PM_{2.5}$ affects mortality through causes beyond LRI.

Third, we can directly compare our main results to quasi-experimental studies from wealthy and middle-income countries that also measured the impact of air quality on infant mortality. We are aware of six quasi-experimental studies^{24–29} that measured the impact of longer-run $PM_{2.5}$ exposure on infant mortality for which we were able to calculate effect sizes in units comparable to ours. As many of these studies reported effect sizes for TSP or PM_{10} , comparing effect sizes requires translating units to $PM_{2.5}$. We used the following conversions: $PM_{2.5} = 0.7PM_{10}$ and $PM_{10} = 0.5TSP$. We note that the results of one study²⁷ are for 0–4 year olds, and baseline mortality rates were not reported in the paper for this study. For another study²⁸, we used the results from table 4, dividing the effect of traffic on mortality by the effect of traffic on $PM_{2.5}$.

Results are shown in Extended Data Fig. 8j. Our results are closest to the previously published results for Mexico City²⁶, with both studies estimating an approximately 9% increase in infant mortality per $10 \mu\text{g m}^{-3}$ change in $PM_{2.5}$. Our results are a little smaller than previously published results published in the US^{24,25} and substantially smaller than recent estimates from urban China, Turkey, and California^{27–29}.

Fourth, there is also growing evidence from both developing and developed countries that $PM_{2.5}$ exposure increases infant mortality risk through causes other than LRI. In particular, multiple studies (including meta-analyses) find that in utero exposure to $PM_{2.5}$ increases the incidence of pre-term birth and low birth weight^{21–23,54–57}, the leading risk factors for neonatal mortality (defined as death in the first 28 days of life). Although the epidemiological evidence leading from $PM_{2.5}$ exposure to adverse fetal outcomes is now substantial, the basic mechanisms are not fully understood and are thought to include pro-inflammatory effects, endocrine effects, neurophysiological effects, metabolic effects, increasing oxidative stress in both the mother and fetus and disruption of oxygen flow^{22,23,58–61}. Studies from both developing and developed countries also directly show impacts of in utero $PM_{2.5}$ exposure on neonatal mortality^{24,25,62}. Although LRIs cause some deaths among neonates, the majority of neonatal mortality is distinct from LRI²⁰. Neonatal mortality thus represents an important channel linking $PM_{2.5}$ exposure and infant mortality that is largely distinct from LRI, and we test for this mechanism in our data below below.

Other work has linked $PM_{2.5}$ exposure to low child height-for-age (that is, stunting)⁶³, an outcome that is primarily reflective of long-term malnutrition⁶⁴ rather than LRI specifically, and which is the leading risk factor for child mortality in Africa. Evidence of a $PM_{2.5}$ –stunting relationship in our data would again be consistent with $PM_{2.5}$ having mortality-relevant health impacts beyond LRI. A few studies also link poor air quality to incidence of diarrhoea^{65,66}. Diarrhoea is a leading cause of infant death in Africa and is mostly considered to be distinct from LRI, even if some conditions, such as poor hygienic conditions, predispose children to both illnesses⁶⁷. Although the mechanisms linking $PM_{2.5}$ exposure and diarrhoeal illness are poorly characterized—for instance it is thought that biological $PM_{2.5}$ could be associated with increased diarrhoea⁶⁶—we nevertheless consider this yet another possible pathway leading from $PM_{2.5}$ exposure to infant mortality that is distinct from LRI, and evidence of this pathway in our data would provide further support for non-LRI effects.

Although the cause of death is unobserved in our data, we use the same empirical strategy as our main analysis to test the extent to which $PM_{2.5}$ levels are related to adverse childhood outcomes that could affect mortality separately from LRI, including neonatal mortality, birth weight, birth size, stunting and diarrhoeal illness. In order to control for differences across mothers (for example, in their ability to estimate birth size), we include mother fixed effects in all specifications so the estimated effects come from comparing outcomes between children born to the same mother at times of different $PM_{2.5}$ concentrations. The point of these analyses is not to isolate the specific causes of death related to $PM_{2.5}$ exposure—this level of ascertainment is not available in DHS, given that DHS does not collect autopsy data—but to provide evidence that $PM_{2.5}$ exposure harms infant health by increasing the risks of several adverse conditions beyond LRI. As shown in Extended Data Fig. 8, we find substantial evidence of non-LRI pathways, as described below.

In utero $PM_{2.5}$ exposure is associated with lower birth weights. We use all observations in our data for which we have birth weight (directly recorded in grams from birth cards and transcribed by DHS enumerators, $n = 215,975$) or birth size recalled by interviewed mothers (five-point scale from ‘very small’ to ‘very large’, $n = 454,444$). Consistent with existing literature on the relationship between PM exposure and birth weight, we find suggestive evidence that higher in utero $PM_{2.5}$ exposure is associated with lower birth weight and birth size, although estimates are imprecise (particularly for the birth card measurements) given the reduced sample.

In utero $PM_{2.5}$ exposure is associated with higher neonatal mortality. Consistent with the epidemiologic evidence that links $PM_{2.5}$ exposure to low birth weight, and

our findings that in utero $PM_{2.5}$ exposure is associated with reduced birth weight and birth size, we find a strong positive link between in utero $PM_{2.5}$ exposure and neonatal mortality (Extended Data Fig. 8c). At the mean exposure in our sample ($25 \mu\text{g m}^{-3}$), we find a similar $PM_{2.5}$ –mortality relationship as in our post-birth results. However, because baseline neonatal mortality rates are less than half infant mortality rates, this effect translates into a larger percentage change. We find a $10 \mu\text{g m}^{-3}$ $PM_{2.5}$ increase associated with an approximately 22% (95% confidence interval, 3–41%) increase in neonatal mortality at mean exposure; the relationship flattens out at higher exposure levels. Overall, the evidence that links in utero $PM_{2.5}$ exposure to birth weight outcomes and neonatal mortality strongly suggests a meaningful non-LRI pathway from $PM_{2.5}$ exposure to infant mortality.

Post-birth exposure to $PM_{2.5}$ reduces child height-for-age among surviving children. Anthropometric measurements are routinely performed on children under five years of age in the majority of DHS surveys. This allows estimation of malnutrition among living children present during the household interviews. Malnutrition is the largest risk factor for mortality in children under five years of age among children in SSA, responsible for over one million deaths of children under-5 from neonatal disorders, diarrhoeal illness, LRI and other common communicable diseases²⁰. Low height-for-age (stunting), in particular, is a reflection of long-term malnutrition. We find that post-birth $PM_{2.5}$ exposure in the first year is strongly associated with reduced height-for-age at the time of survey (Extended Data Fig. 8d). To the extent that high $PM_{2.5}$ represents unhealthy environments for child growth, it could increase the risk of stunting. Stunting, in turn, represents another pathway linking $PM_{2.5}$ exposure with infant mortality that is unrelated to LRI. We note that the sample for this analysis only includes children who survived to the date of the interview, and thus undercounts children who died before the survey and who may be smaller in size; this may lead us to underestimate the ‘true’ effect of $PM_{2.5}$ on stunting in this analysis.

Post-birth $PM_{2.5}$ exposure is associated with increase diarrhoeal incidence among surviving children. Interviewed mothers are asked whether or not each living child born in the past 3–5 years had diarrhoea in the two weeks before the survey. We again find a positive relationship between post-birth $PM_{2.5}$ exposure in the first year and the probability of experiencing diarrhoea within the two weeks preceding the survey (Extended Data Fig. 8e). We interpret this as additional evidence of a potential pathway beyond LRI, and one that should be investigated further.

Post-birth $PM_{2.5}$ exposure has no effect on child sex, the likelihood of a multiple birth or bed net usage. These measures are the three most easily observable child-level variables in our data (Extended Data Fig. 8f–h). These are placebo tests to ensure that our main effects are not spurious and that $PM_{2.5}$ exposure is uncorrelated with variables that should not be affected by air quality.

$PM_{2.5}$ exposure 13–24 months after birth does not predict mortality in the first 12 months after birth. As another placebo test, we confirm that exposure after the first birthday of a child does not affect the probability of dying before the first birthday (Extended Data Fig. 8i). This again is evidence that our main effects are not spurious.

The balance of evidence thus strongly suggests that $PM_{2.5}$ exposure is associated with infant survival through channels other than respiratory infection, and lends plausibility to our main results. For instance, the GBD estimates that LRI (12%), diarrhoeal illnesses (11%) and neonatal disorders (27%) make up a combined 50% of all deaths of children under five in SSA²⁰. Given evidence that $PM_{2.5}$ is associated with all of these channels, and the fact that our estimates are even smaller than comparable estimates from the US, our estimate that approximately 20% of infant mortality could be attributed to $PM_{2.5}$ exposure is plausible. Nevertheless, we view prospective epidemiological studies that measure both exposures and intermediary outcomes as critical in building a more complete understanding of causal pathways going forward; such studies would shed critical light on the large overall estimates that we provide here.

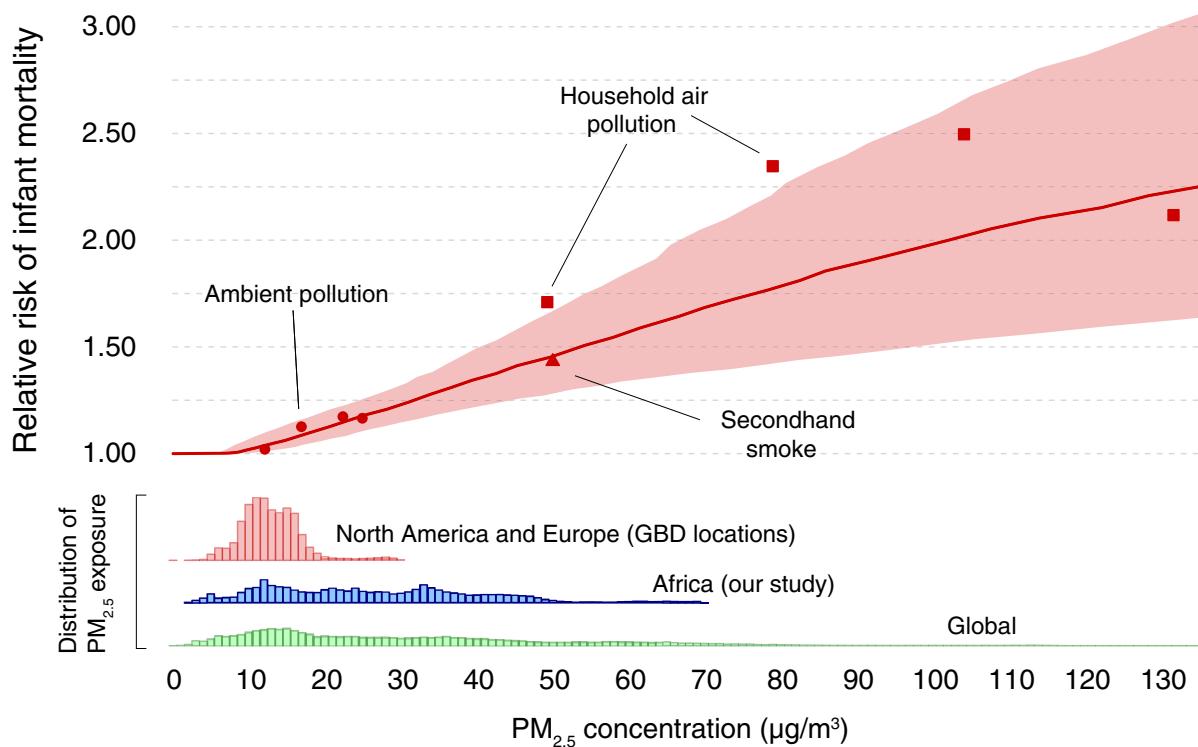
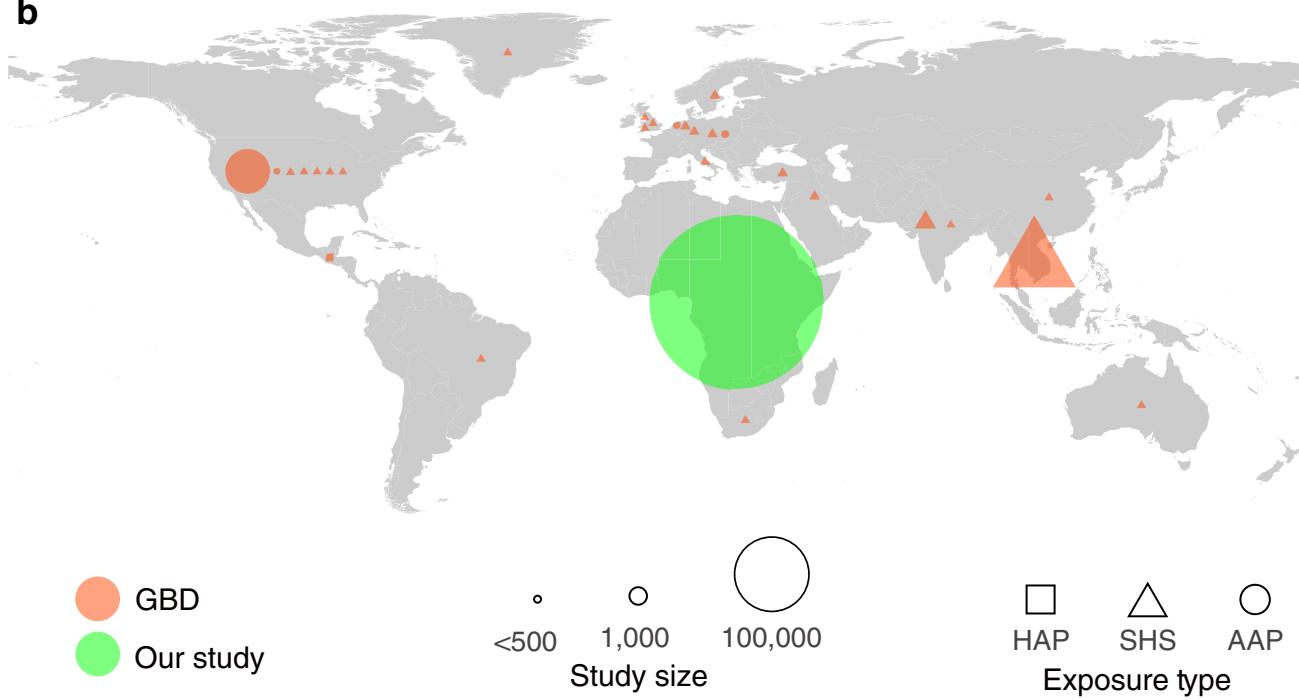
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. All codes that support the findings of this study are available at <https://purl.stanford.edu/qt056zr6479>.

Data availability. All data and code that support the findings of this study are available at <https://purl.stanford.edu/qt056zr6479>.

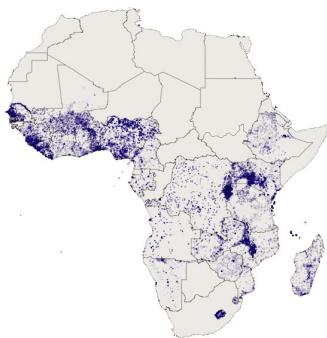
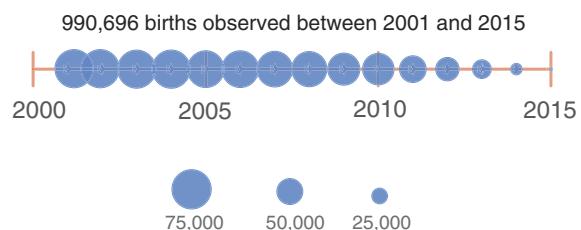
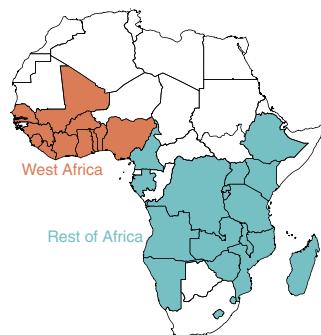
31. ICF. The DHS program, Data. <http://dhsprogram.com/data> (1998).
32. Bendavid, E. Is health aid reaching the poor? Analysis of household data from aid recipient countries. *PLoS ONE* **9**, e84025 (2014).
33. Lee, H. J., Coull, B. A., Bell, M. L. & Koutrakis, P. Use of satellite-based aerosol optical depth and spatial clustering to predict ambient $PM_{2.5}$ concentrations. *Environ. Res.* **118**, 8–15 (2012).
34. Hyder, A. et al. $PM_{2.5}$ exposure and birth outcomes: use of satellite- and monitor-based data. *Epidemiology* **25**, 58–67 (2014).
35. Crouse, D. L. et al. A new method to jointly estimate the mortality risk of long-term exposure to fine particulate matter and its components. *Sci. Rep.* **6**, 18916 (2016).

36. Burke, M., Gong, E. & Jones, K. Income shocks and HIV in Africa. *Econ. J.* **125**, 1157–1189 (2015).
37. Funk, C. et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* **2**, 150066 (2015).
38. LP DAAC. MODIS/aqua land surface temperature and emissivity (LST/E), version 6. USGS Earth Resources Observation And Science Center. <https://lpdaac.usgs.gov> (NASA, 2016).
39. Baird, S., Friedman, J. & Schady, N. Aggregate income shocks and infant mortality in the developing world. *Rev. Econ. Stat.* **93**, 847–856 (2011).
40. Tatem, A. J. et al. Millennium development health metrics: where do Africa's children and women of childbearing age live? *Popul. Health Metr.* **11**, 11 (2013).
41. Rehman, I., Ahmed, T., Praveen, P., Kar, A. & Ramanathan, V. Black carbon emissions from biomass and fossil fuels in rural India. *Atmos. Chem. Phys.* **11**, 7289–7299 (2011).
42. Anenberg, S. C., Horowitz, L. W., Tong, D. Q. & West, J. J. An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling. *Environ. Health Perspect.* **118**, 1189–1195 (2010).
43. Isen, A., Rossin-Slater, M. & Walker, W. R. Every breath you take, every dollar you'll make: the long-term consequences of the clean air act of 1970. *J. Polit. Econ.* **125**, 848–902 (2017).
44. Wilson, W. E. & Suh, H. H. Fine particles and coarse particles: concentration relationships relevant to epidemiologic studies. *J. Air Waste Manag. Assoc.* **47**, 1238–1249 (1997).
45. Brook, J. R., Dann, T. F. & Burnett, R. T. The relationship among TSP, PM₁₀, PM_{2.5}, and inorganic constituents of atmospheric particulate matter at multiple Canadian locations. *J. Air Waste Manag. Assoc.* **47**, 2–19 (1997).
46. Dionisio, K. L. et al. Measuring the exposure of infants and children to indoor air pollution from biomass fuels in the Gambia. *Indoor Air* **18**, 317–327 (2008).
47. Ezzati, M., Saleh, H. & Kammen, D. M. The contributions of emissions and spatial microenvironments to exposure to indoor air pollution from biomass combustion in Kenya. *Environ. Health Perspect.* **108**, 833–839 (2000).
48. Kilabuko, J. H., Matsuki, H. & Nakai, S. Air quality and acute respiratory illness in biomass fuel using homes in Bagamoyo, Tanzania. *Int. J. Environ. Res. Public Health* **4**, 39–44 (2007).
49. Yamamoto, S. S., Louis, V. R., Sié, A. & Sauerborn, R. Biomass smoke in Burkina Faso: what is the relationship between particulate matter, carbon monoxide, and kitchen characteristics? *Environ. Sci. Pollut. Res.* **21**, 2581–2591 (2014).
50. Fischer, S. L. & Koshland, C. P. Daily and peak 1 h indoor air pollution and driving factors in a rural Chinese village. *Environ. Sci. Technol.* **41**, 3121–3126 (2007).
51. Ni, K. et al. Seasonal variation in outdoor, indoor, and personal air pollution exposures of women using wood stoves in the Tibetan Plateau: baseline assessment for an energy intervention study. *Environ. Int.* **94**, 449–457 (2016).
52. Parikh, J., Balakrishnan, K., Laxmi, V. & Biswas, H. Exposure from cooking with biofuels: pollution monitoring and analysis for rural Tamil Nadu, India. *Energy* **26**, 949–962 (2001).
53. Cyrys, J., Pitz, M., Bischof, W., Wichmann, H.-E. & Heinrich, J. Relationship between indoor and outdoor levels of fine particle mass, particle number concentrations and black smoke under different ventilation conditions. *J. Expo. Anal. Environ. Epidemiol.* **14**, 275–283 (2004).
54. Pedersen, M. et al. Ambient air pollution and low birthweight: a European cohort study (ESCAPE). *Lancet Respir. Med.* **1**, 695–704 (2013).
55. Chang, H. H., Warren, J. L., Darrow, L. A., Reich, B. J. & Waller, L. A. Assessment of critical exposure and outcome windows in time-to-event analysis with application to air pollution and preterm birth study. *Biostatistics* **16**, 509–521 (2015).
56. Li, X. et al. Association between ambient fine particulate matter and preterm birth or term low birth weight: an updated systematic review and meta-analysis. *Environ. Pollut.* **227**, 596–605 (2017).
57. Blum, J. L., Chen, L.-C. & Zelikoff, J. T. Exposure to ambient particulate matter during specific gestational periods produces adverse obstetric consequences in mice. *Environ. Health Perspect.* **125**, 077020 (2017).
58. Kampa, M. & Castanas, E. Human health effects of air pollution. *Environ. Pollut.* **151**, 362–367 (2008).
59. Slama, R. et al. Meeting report: atmospheric pollution and human reproduction. *Environ. Health Perspect.* **116**, 791–798 (2008).
60. Jerrett, M. et al. Traffic-related air pollution and obesity formation in children: a longitudinal, multilevel analysis. *Environ. Health* **13**, 49 (2014).
61. Suades-González, E., Gascon, M., Guxens, M. & Sunyer, J. Air pollution and neuropsychological development: a review of the latest evidence. *Endocrinology* **156**, 3473–3482 (2015).
62. Jayachandran, S. Air quality and early-life mortality: evidence from Indonesia's wildfires. *J. Hum. Ressour.* **44**, 916–954 (2009).
63. Tielsch, J. M. et al. Exposure to indoor biomass fuel and tobacco smoke and risk of adverse reproductive outcomes, mortality, respiratory morbidity and growth among newborn infants in south India. *Int. J. Epidemiol.* **38**, 1351–1363 (2009).
64. Black, R. E. et al. Maternal and child undernutrition: global and regional exposures and health consequences. *Lancet* **371**, 243–260 (2008).
65. Malilay, J., Mariana, G. R., Ramirez Vanegas, A., Non, E. & Sinks, T. Public health surveillance after a volcanic eruption: lessons from Cerro Negro, Nicaragua, 1992. *Bull. Pan. Am. Health. Organ.* **30**, 218–226 (1996).
66. Menetrez, M. Y. et al. An evaluation of indoor and outdoor biological particulate matter. *Atmos. Environ.* **43**, 5476–5483 (2009).
67. Luby, S. P. et al. Effect of handwashing on child health: a randomised controlled trial. *Lancet* **366**, 225–233 (2005).

a**b**

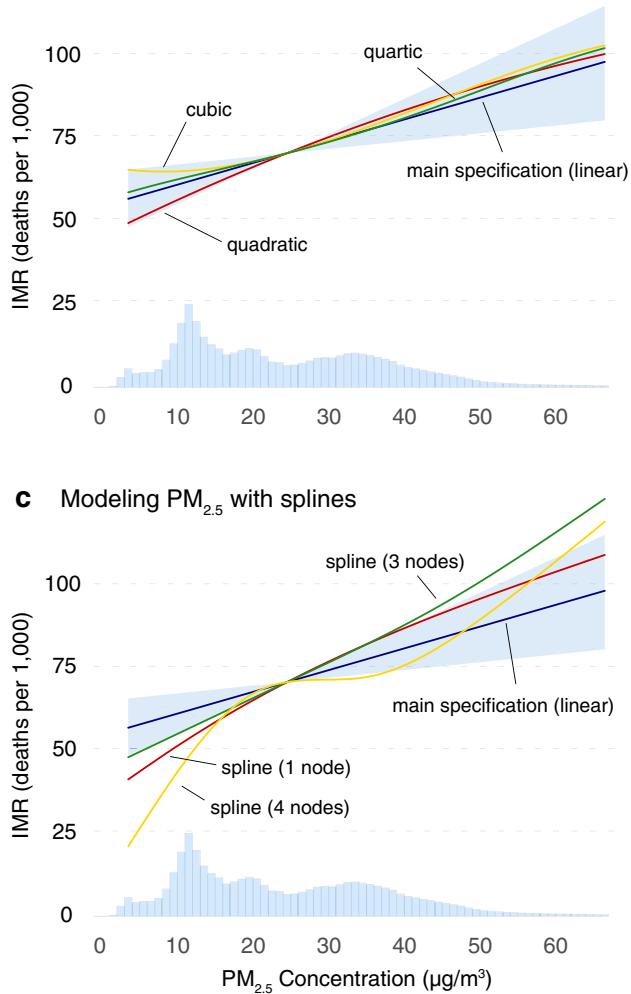
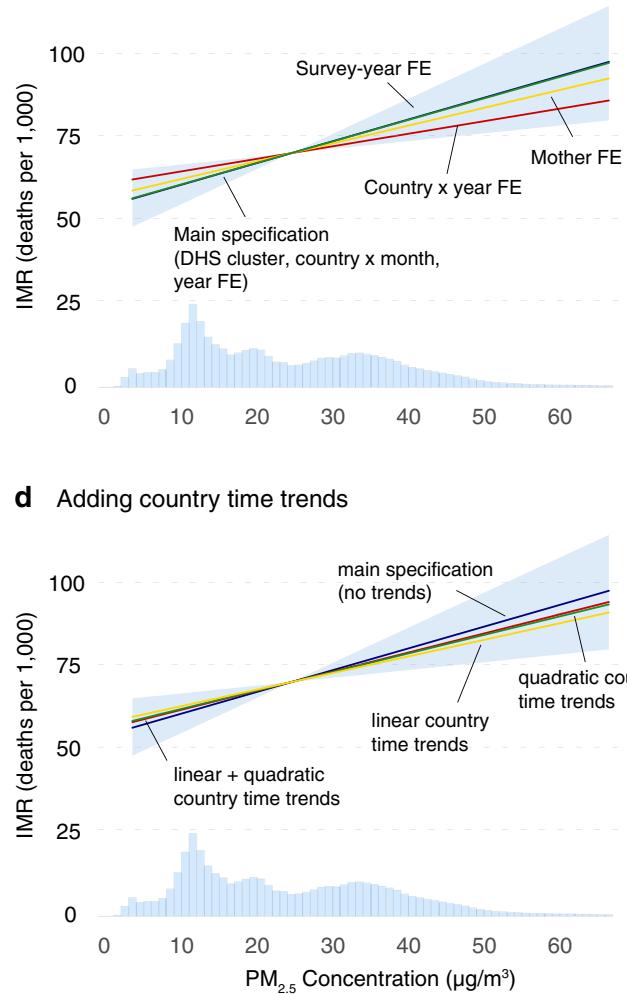
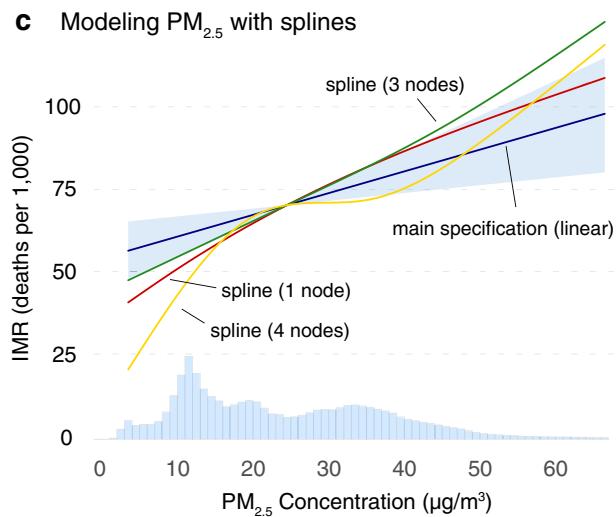
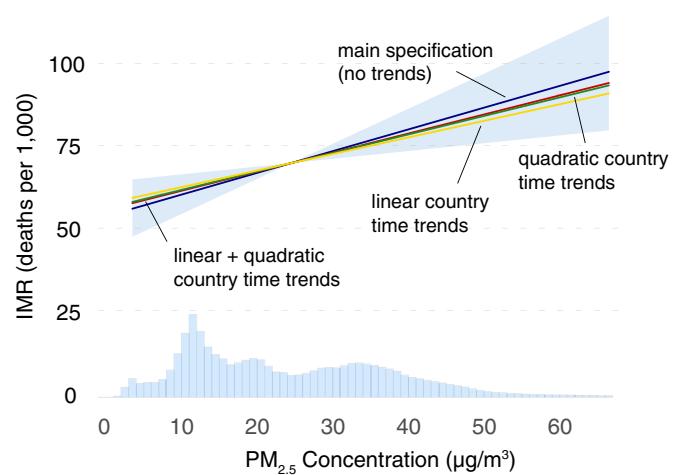
Extended Data Fig. 1 | Integrated exposure risk curve estimated by the GBD project. Data were obtained from a previous study⁴. **a**, Relative risk curve representing the risk from acute lower respiratory infections in infants (obtained from figure 2 of Burnett et al.⁴). The curve combines point estimates from ambient air pollution (AAP) studies, indoor air pollution (HAP) studies and second-hand smoking (SHS) studies to derive risk responses across the PM_{2.5} exposure distribution. The histograms show the share of population exposed to different long-run (15-year average) ambient PM_{2.5} concentrations in North American and Europe where most GBD studies took place, in SSA countries in our sample,

and globally. In total, 49% of the overall population in Africa, and 51% globally, live in areas with ambient pollution concentrations exceeding the maximum ambient PM_{2.5} concentration from the GBD study (25 $\mu\text{g m}^{-3}$). **b**, Most studies used to estimate the GBD integrated exposure response⁴ were carried out in North America or Europe, with the exception of a household air pollution study in Guatemala and second-hand smoking studies in Vietnam, India and South Africa. Median sample size (depicted by marker size in the plot) across these studies is $n = 1,250$. Country outlines were obtained from Global Administrative Areas, version 2.0³⁰.

a DHS cluster locations**b** Number of births observed by year**c** Study regions

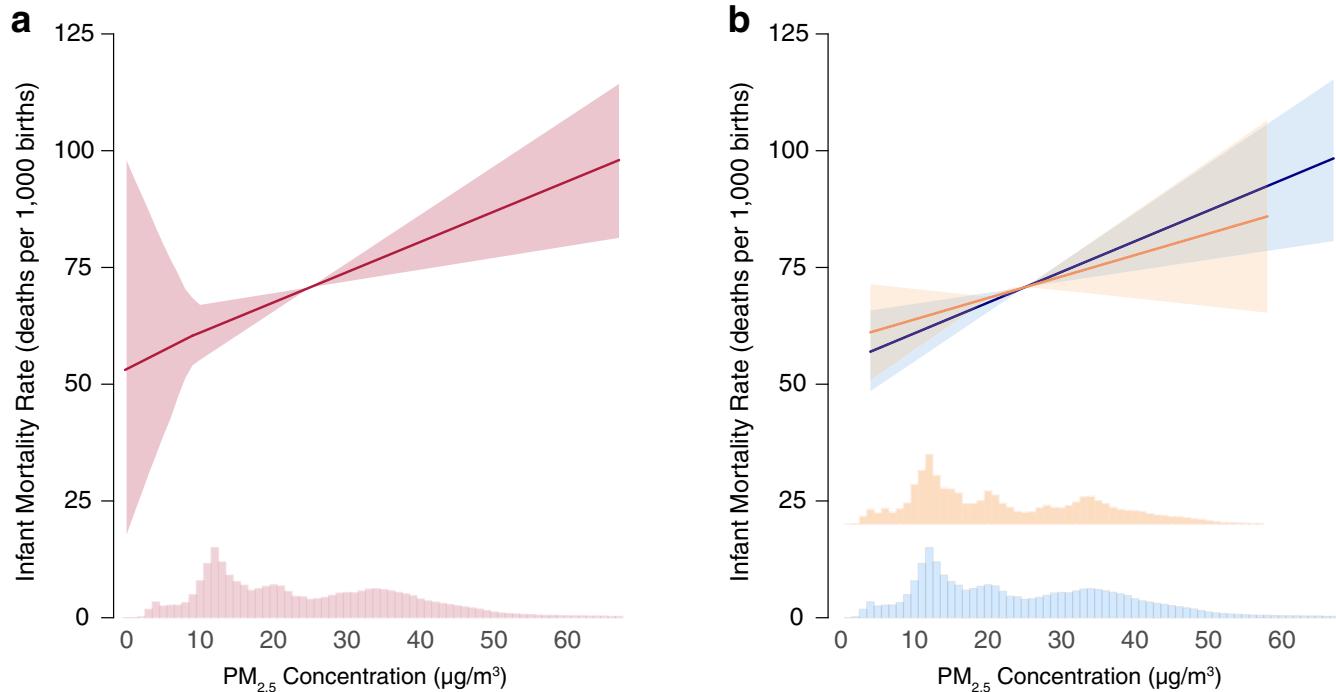
Extended Data Fig. 2 | Overview of birth data from DHS surveys and study regions in Africa. **a**, Location of DHS clusters included in our sample. **b**, The number of births observed in each year in our sample. More births are observed in earlier years because births are recalled in the surveys so each new survey round potentially adds births from all previous years. **c**, Regional categorization of countries, for regional analysis in Fig. 2c. Sample countries assigned to West Africa region are Benin,

Burkina Faso, Ivory Coast, Ghana, Guinea, Liberia, Mali, Nigeria, Senegal, Sierra Leone and Togo. Sample countries assigned to 'rest of Africa' are Angola, Burundi, Cameroon, Comoros, DRC, Ethiopia, Gabon, Kenya, Lesotho, Madagascar, Malawi, Mozambique, Namibia, Rwanda, Swaziland, Uganda, Zambia and Zimbabwe. Country outlines were obtained from Global Administrative Areas, version 2.0³⁰.

a Modeling PM_{2.5} with polynomials**b Different fixed effects****c Modeling PM_{2.5} with splines****d Adding country time trends**

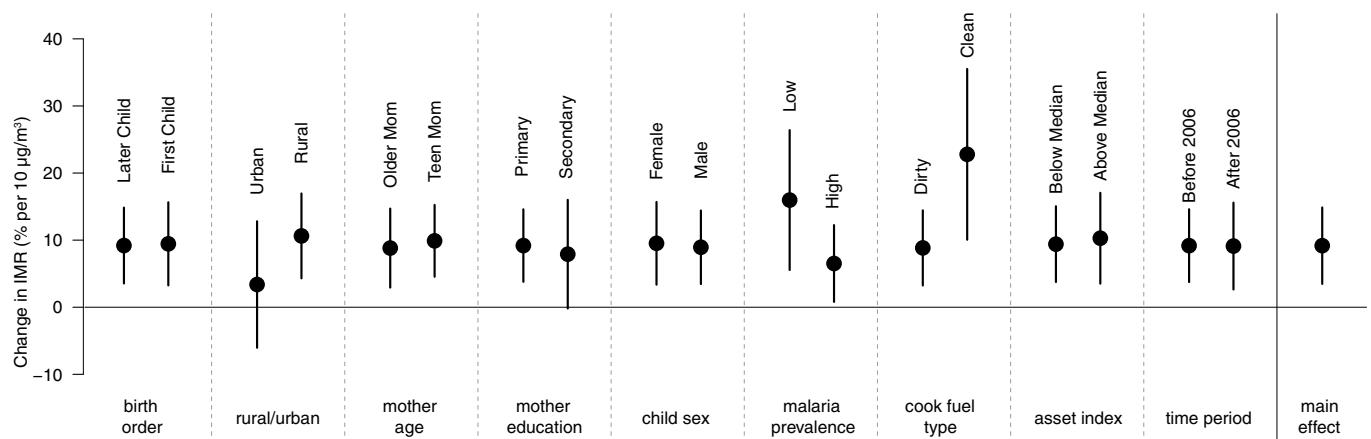
Extended Data Fig. 3 | Effect of post-birth PM_{2.5} exposure is robust under different regression models. Estimated responses under higher-order polynomials (a), different specifications of the fixed effects (b), restricted cubic spline functions of PM_{2.5} (c) and additional time

controls (d). In each panel, the blue line and shaded region indicate the estimated baseline response shown in Fig. 2a and the bootstrapped 95% confidence interval. Splines in c have knots at $10 \mu\text{g m}^{-3}$ (single knot spline) or evenly spaced knots (three- and four-knot splines).



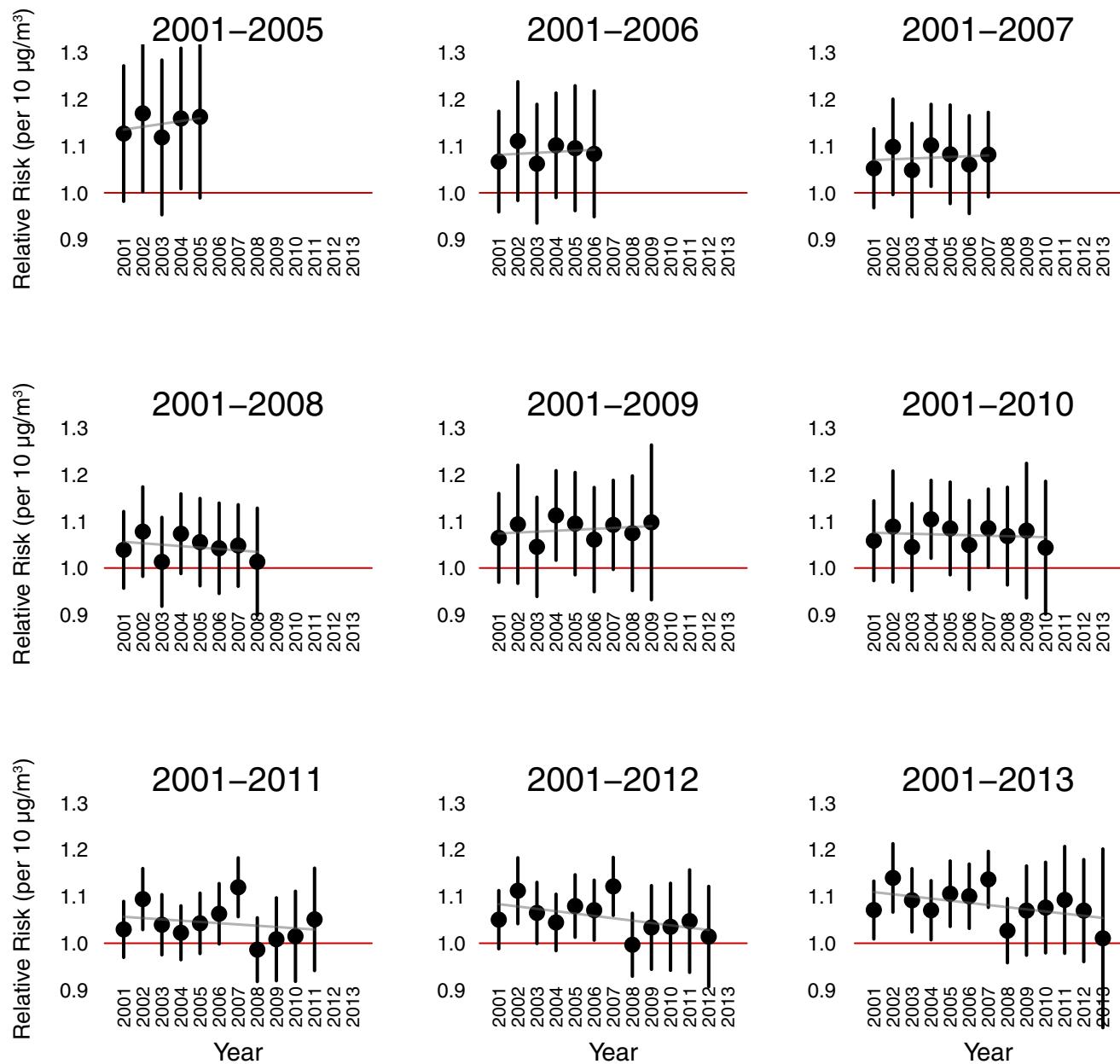
Extended Data Fig. 4 | Piecewise linear and cross-sectional relationships between post-birth $\text{PM}_{2.5}$ exposure and infant mortality.
a, Piecewise linear estimates of the effect of $\text{PM}_{2.5}$ exposure below and above the WHO $\text{PM}_{2.5}$ guideline of $10 \mu\text{g m}^{-3}$. Shaded regions represent bootstrapped 95% confidence intervals. Slopes above and below the $10 \mu\text{g m}^{-3}$ threshold are very similar, although confidence intervals are wider below the threshold due to smaller sample sizes. **b**, Cross-sectional and panel models give similar estimated effects of post-birth $\text{PM}_{2.5}$

exposure on infant mortality. Blue line shows baseline panel model, orange line shows a cross-sectional model that relates cluster-average mortality to cluster-average $\text{PM}_{2.5}$ exposure. Each response function is centred at sample median exposure ($25 \mu\text{g m}^{-3}$). Histograms at the bottom show counts of exposure at different $\text{PM}_{2.5}$ levels, for the panel sample (blue) and cross-sectional sample (orange); cross-sectional exposures are slightly narrower given that year-to-year variation has been averaged out.



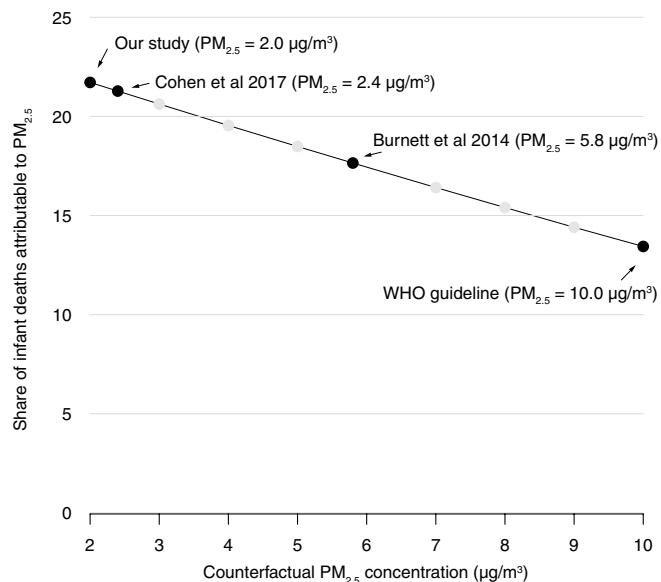
Extended Data Fig. 5 | Heterogeneous effects of post-birth PM_{2.5} exposure. Effects are estimated by interacting a dummy for each modifying variable with linear PM_{2.5}, and are measured as the percentage change in infant mortality per 10 µg m⁻³ increase in PM_{2.5} exposure,

relative to baseline mortality rates in each subgroup. Circles indicate point estimates, and whiskers the 95% confidence interval on the point estimate. The last column shows the baseline estimate from the full (uninteracted) linear model.

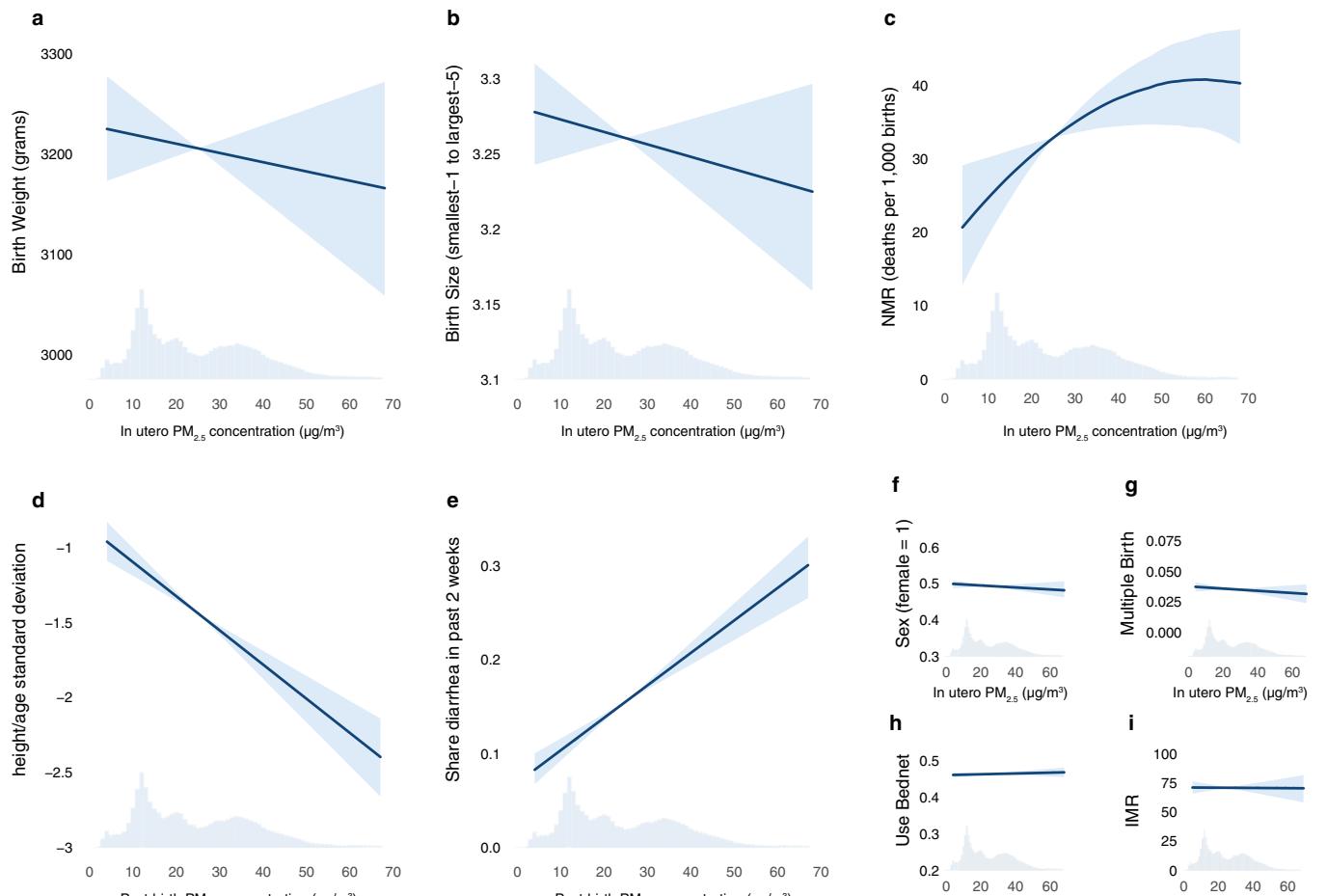


Extended Data Fig. 6 | Linear effect of post-birth PM_{2.5} exposure by year for different time periods. Panels are the same as Fig. 2e but replicated for different time periods, showing effects in each year independently. Circles indicate point estimates, and whiskers the 95%

confidence interval on the point estimate. For each time period 2001 – year t , the sample was restricted to births between 2001 and year t and to surveys that were conducted after year t . These steps help to approximate a consistent geographical sample across the time periods.



Extended Data Fig. 7 | Effect of different assumed counterfactual PM_{2.5} levels on the estimated share of infant deaths attributable to PM_{2.5}. Each point represents the same calculation described in the Methods, under different counterfactual minimum PM_{2.5} exposure levels. Data are from Cohen et al.², Burnett et al.⁴ and the WHO guidelines¹⁶.



Study	Period	Location	Baseline mortality rate (per 100,000)	Effect size (% increase in infant mortality rate per 10 $\mu\text{g}/\text{m}^3$ $\text{PM}_{2.5}$)
Arceo, Hanna, Oliva 2015 ²⁶	1997-2006	Mexico City	1987	8.8
Chay and Greenstone 2003a ²⁴	1978-1982	US	1226	11.0
Chay and Greenstone 2003b ²⁵	1969-1974	US	1899	11.3
Cesur et al 2016 ²⁹	2001-2011	Turkey	900	21.6
He et al 2016 ²⁷	2006-2010	China (urban)	NA	27.1
Knittel et al 2016 ²⁸	2002-2007	California	280	34.3
<i>This study</i>	2001-2015	Africa	7076	9.2

Extended Data Fig. 8 | Effect of $\text{PM}_{2.5}$ on non-respiratory mortality and mortality risk factors. **a–c**, Effect of in utero $\text{PM}_{2.5}$ exposure on low birth weight, low birth size as reported by mothers on a scale from 1 to 5, and neonatal mortality (NMR). **d, e**, Effect of post-birth $\text{PM}_{2.5}$ exposure on height-for-age and diarrhoeal incidence for living children. In each case, higher $\text{PM}_{2.5}$ concentrations worsen health outcomes. **f–h**, Placebo tests that relate $\text{PM}_{2.5}$ exposures to child outcomes that should be unaffected:

child sex, whether child was born in a multiple birth, and child's use of a bed net. **i**, $\text{PM}_{2.5}$ exposure in the 13–24 months after birth has no effect on mortality in the first 12 months after birth. Shaded regions represent bootstrapped 95% confidence intervals in each panel. **j**, Estimates of the effect of $\text{PM}_{2.5}$ on all-cause infant mortality from published quasi-experimental studies^{24–29}, expressed as the percentage change in the infant mortality rate per $10 \mu\text{g}/\text{m}^3$ increase in $\text{PM}_{2.5}$.

Extended Data Table 1 | Regression results for main specification and for subsample of households with asset data

	a					
	Linear Model - full sample			Quadratic Model- full sample		
	(1)	(2)	(3)	(4)	(5)	(6)
PM _{2.5}	0.000656*** (0.0001938)	0.0006412*** (0.0001984)	0.0006492*** (0.0002016)	0.0010768** (0.0005449)	0.0010592* (0.0005523)	0.0011407** (0.0005403)
PM _{2.5} -squared				-0.000003986 (0.000005000)	-0.000003953 (0.000005037)	-0.000004648 (0.000005032)
Observations	990,696	990,696	990,696	990,696	990,696	990,696
Controls:						
Temp & Rainfall	No	Yes	Yes	No	Yes	Yes
Individual Covariates	No	No	Yes	No	No	Yes
Asset Wealth Index	No	No	No	No	No	No
Regression Weights	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

	b			
	Linear Model - wealth sample		Quadratic Model - wealth sample	
	(1)	(2)	(3)	(4)
PM _{2.5}	0.0006912*** (0.0002086)	0.0006877 *** (0.0002089)	0.0015868 *** (0.0005826)	0.0015762 *** (0.0005824)
PM _{2.5} -squared			-0.000008 (0.000005)	-0.000008 (0.000005)
Observations	833,001	833,001	833,001	833,001
Controls:				
Asset Wealth Index	No	Yes	No	Yes
Temp & Rainfall	Yes	Yes	Yes	Yes
Individual Covariates	Yes	Yes	Yes	Yes
Regression Weights	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes

a, Full sample. **b**, Subsample of households with asset data. Mortality is modelled as either a linear or quadratic function of PM_{2.5} (results from additional specifications modelling mortality as a more flexible nonlinear function of PM_{2.5} are shown in Extended Data Fig. 3). The outcome measure is a binary variable equal to one if the child did not survive until 12 months of age. The mean of the outcome variable = 0.071. Standard errors are in parentheses and are clustered at the DHS Cluster level, and asterisks denote statistical significance (two-sided): * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

Extended Data Table 2 | Understanding potential bias from unobserved indoor air pollution exposure

a

	$0 < \delta < 1$ mass outflow > mass inflow	$\delta = 1$ mass outflow = mass inflow	$\delta > 1$ mass inflow > mass outflow
$\delta_0 = 0$	[1] $PM_{in} < PM_{out}$ $\hat{\alpha}$ biased down	[2] $PM_{in} = PM_{out}$ $\hat{\alpha}$ unbiased	[3] $PM_{in} > PM_{out}$ $\hat{\alpha}$ biased up
\Rightarrow No Indoor Source			
$\delta_0 > 0$	[4] $PM_{in} \geq PM_{out}$ $\hat{\alpha}$ biased down	[5] $PM_{in} > PM_{out}$ $\hat{\alpha}$ unbiased	[6] $PM_{in} > PM_{out}$ $\hat{\alpha}$ biased up
\Rightarrow Indoor Source			

b

	Contemporaneous Ambient PM _{2.5} Concentration					
	All DHS Clusters			Clusters w/ Clean Cooking $\neq 0$		
	(1) Full	(2) Urban	(3) Rural	(4) Full	(5) Urban	(6) Rural
Clean cooking fuel penetration rate	-3.177*** (0.347)	-4.931*** (0.403)	-2.687** (0.951)	-8.563*** (0.846)	-9.869*** (0.981)	-6.885*** (1.964)
Constant	22.94*** (0.0866)	24.51*** (0.170)	22.39*** (0.100)	27.52*** (0.669)	28.84*** (0.807)	25.35*** (1.239)
N	25483	7789	17694	2444	1952	492
R ²	0.00307	0.0180	0.000417	0.0414	0.0498	0.0249
RMSE	13.23	13.16	13.22	12.96	12.86	13.28

a. Potential bias at the household level in the estimated effect of PM_{2.5} exposure on infant health $\hat{\alpha}$ as a function of the relationship between indoor and outdoor pollution exposure. Cells show the expected relative magnitudes of time-averaged differences in PM_{in} versus PM_{out} for all combinations of δ_0 and δ . **b.** Relationship between ambient PM_{2.5} and fraction of households using clean cooking fuels at the DHS cluster level. Standard errors are in parentheses, and asterisk denote statistical significance (two-sided): * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$. RMSE, root mean squared error.

Species-specific activity of antibacterial drug combinations

Ana Rita Brochado¹, Anja Telzerow¹, Jacob Bobonis¹, Manuel Banzhaf^{1,11}, André Mateus¹, Joel Selkirk¹, Emily Huth², Stefan Bassler¹, Jordi Zamarreño Beas³, Matylda Zietek¹, Natalie Ng⁴, Sunniva Foerster⁵, Benjamin Ezraty³, Béatrice Py³, Frédéric Barras^{3,6}, Mikhail M. Savitski¹, Peer Bork^{7,8,9,10}, Stephan Göttig² & Athanasios Typas^{1,7*}

The spread of antimicrobial resistance has become a serious public health concern, making once-treatable diseases deadly again and undermining the achievements of modern medicine^{1,2}. Drug combinations can help to fight multi-drug-resistant bacterial infections, yet they are largely unexplored and rarely used in clinics. Here we profile almost 3,000 dose-resolved combinations of antibiotics, human-targeted drugs and food additives in six strains from three Gram-negative pathogens—*Escherichia coli*, *Salmonella enterica* serovar Typhimurium and *Pseudomonas aeruginosa*—to identify general principles for antibacterial drug combinations and understand their potential. Despite the phylogenetic relatedness of the three species, more than 70% of the drug–drug interactions that we detected are species-specific and 20% display strain specificity, revealing a large potential for narrow-spectrum therapies. Overall, antagonisms are more common than synergies and occur almost exclusively between drugs that target different cellular processes, whereas synergies are more conserved and are enriched in drugs that target the same process. We provide mechanistic insights into this dichotomy and further dissect the interactions of the food additive vanillin. Finally, we demonstrate that several synergies are effective against multi-drug-resistant clinical isolates *in vitro* and during infections of the larvae of the greater wax moth *Galleria mellonella*, with one reverting resistance to the last-resort antibiotic colistin.

To study the characteristics and conservation of drug–drug interactions in bacteria, we selected three gamma-proteobacterial species, *E. coli*, *S. Typhimurium* and *P. aeruginosa*, all of which belong to the highest-risk group for antibiotic resistance³. We used model laboratory strains rather than multi-drug-resistant (MDR) isolates to derive general principles behind drug–drug interactions without being confounded by horizontally transferred antibiotic resistance elements, and to facilitate follow-up experiments and comparisons with previous and future results. To further assess whether drug responses vary between strains of the same species, we included two strains per species (Extended Data Fig. 1a), probing each in up to 79 compounds alone and in pairwise combinations. The compounds comprised 59% antibiotics (covering all major classes), 23% human-targeted drugs and food additives—most of which have reported antibacterial and/or adjuvant activity^{4,5}—and 18% other compounds with known bacterial targets or genotoxic effects, such as proton motive force inhibitors or oxidative damage agents, owing to their potential relevance for antibiotic activity and/or uptake^{6,7} (Extended Data Fig. 1a, Supplementary Table 1). We profiled up to 2,883 pairwise drug combinations in each of the 6 strains (17,050 combinations in total). We assessed each drug combination in a 4 × 4 tailored-dose matrix (Methods, Supplementary Table 1), used optical density as growth readout and calculated fitness as the growth ratio between drug-treated and untreated cells (Extended Data Figs. 1, 2,

Methods). All experiments were done at least twice, and on average four times, with high replicate correlation (average Pearson correlation = 0.93, Extended Data Fig. 3a, b).

We quantified all drug–drug interactions using the Bliss independence model (Extended Data Fig. 1b, Methods). Consistent with its null hypothesis, interaction scores were zero-centred for all species (Extended Data Fig. 3c). From all the scores (ε) obtained per combination (4 × 4 dose matrix), we derived a single interaction score ($\bar{\varepsilon}$) that ranged from –1 to 1 (Methods). Synergies and antagonisms were considered significant if $P < 0.05$ (Benjamini–Hochberg corrected, 10,000 repetitions of a two-sided Wilcoxon rank-sum test). Strong interactions had an additional effect size requirement for $|\bar{\varepsilon}| > 0.1$, whereas weak interactions could satisfy the effect-size threshold for one of the two strains of the same species but be slightly below the threshold for the other ($|\bar{\varepsilon}| > 0.06$, Methods). In total, we detected 19% interactions (synergies and antagonisms combined) for *E. coli*, 16% for *S. Typhimurium* and 11% for *P. aeruginosa* (Supplementary Table 2). These hit rates are between the >70% hit rate for 21 antibiotics previously tested in *E. coli*⁸ and the <2% hit rate for a larger set of combinations previously tested in a number of different fungal species⁹. Discrepancies are likely due to: (i) drug selection biases, (ii) single-drug concentrations used in previous studies, which increase false-negative and -positive rates, and (iii) different strategies of data analysis. For example, we observed that drugs that lack antibacterial activity engage in fewer interactions (Extended Data Fig. 3e): the previous study in fungi⁹ screened pairwise combinations of 6 antifungals with 3,600 drugs, most of which had no antifungal activity—probably explaining the low number of interactions detected—and the study in *E. coli*⁸ profiled only bioactive antibiotics. Out of the 79 drugs tested here, all had at least 1 interaction, and a median of 5–13 interactions, in the different strains (Extended Data Fig. 3f).

Because, to our knowledge, drug combinations have not previously been systematically probed in bacteria, we lacked a ground truth for benchmarking our dataset. To overcome this limitation, we selected 242 combinations and created a validation set using higher-precision 8 × 8 checkerboard assays (Extended Data Fig. 4a, b, Supplementary Table 3, Methods). We used this validation set both to assess the performance of our interaction identification approach and to benchmark our screen (Extended Data Fig. 4c, d). Overall, we had a precision and recall of 91% and 74%, respectively. The slightly lower recall can be partially explained by the larger coverage of drug concentration range in the validation experiments, which improves our ability to detect interactions (Extended Data Fig. 5). We confirmed 90% of all the weak interactions that we probed in the validation set ($n = 46$, Extended Data Fig. 6, Supplementary Table 3), which supports the rationale of our interaction identification approach. Indeed, including weak interactions in our

¹European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. ²Institute of Medical Microbiology and Infection Control, Hospital of Goethe University, Frankfurt am Main, Germany. ³Laboratoire de Chimie Bactérienne, Institut de Microbiologie de la Méditerranée, CNRS UMR 7283, Aix-Marseille Université, Marseille, France. ⁴Department of Bioengineering, Stanford University, Stanford, CA, USA. ⁵Institute of Social & Preventive Medicine, Institute of Infectious Diseases, University of Bern, Bern, Switzerland. ⁶Institut Pasteur, Paris, France. ⁷European Molecular Biology Laboratory, Structural & Computational Biology Unit, Heidelberg, Germany. ⁸Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany. ⁹Molecular Medicine Partnership Unit, Heidelberg, Germany. ¹⁰Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany. ¹¹Present address: Institute of Microbiology & Infection, School of Biosciences, University of Birmingham, Birmingham, UK. *e-mail: typas@embl.de

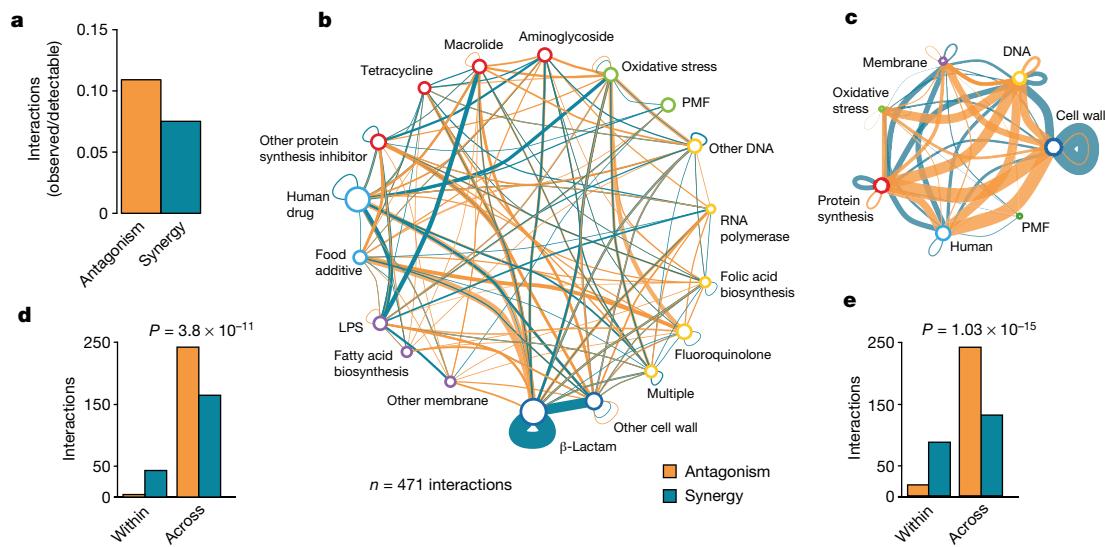


Fig. 1 | Principles of drug–drug interaction networks. **a**, Antagonism is more prevalent than synergy. Fraction of observed over detectable interactions for the six strains. We detect more antagonistic (1,354) than synergistic (1,230) interactions, although our ability to detect antagonisms is lower than our ability to detect synergies (12,778 versus 16,920 combinations). **b, c**, Drug–drug interaction networks in *E. coli*. Nodes represent either drug categories (**b**) or drugs grouped according to the general cellular process that they target (**c**). Node colours represent general cellular processes: blue, cell wall; yellow, DNA; red, protein synthesis; teal, human-targeted or food additive; lilac, membrane; green, oxidative stress or protein motor force (PMF). Node size reflects the number of drugs

hits increases the recall (Extended Data Fig. 4d). For a handful of the synergies observed between antibiotics of the same class (β -lactams), we confirmed the interactions using the Loewe additivity model (Extended Data Fig. 4e), which is more suitable for assessing interactions between drugs with the same target.

Overall, we detected 1,354 antagonistic and 1,230 synergistic interactions. Although this suggests that the two occur with similar frequencies, antagonisms are nearly 50% more prevalent than synergies after correcting for our ability to detect both types of interactions (Fig. 1a). This is because we can detect antagonisms only for 75% of combinations (when at least one drug inhibits growth; Extended Data Fig. 3d, Methods), whereas synergies are detectable for nearly all combinations. A higher prevalence of antagonisms has also been reported for antifungals¹⁰.

Notably, antagonisms and synergies exhibited a clear dichotomy in our data. Antagonism occurred almost exclusively between drugs that target different cellular processes, whereas synergies were also abundant for drugs of the same class or that target the same process (Fig. 1b–e, Extended Data Fig. 7). Mechanistically, antagonism can be explained by interactions at the drug-target level, as the two inhibitors can help the cell to buffer the distinct processes that are perturbed. DNA and protein synthesis inhibitors act this way in bacteria¹¹ (Fig. 1b). Consistent with this being a broader phenomenon, in genome-wide genetic interactions studies in yeast, alleviating interactions (antagonisms) are enriched between essential genes (the targets of anti-infectives), which are part of different functional processes¹². However, antagonism can also occur at the level of intracellular drug concentrations (Extended Data Fig. 8a). We tested 16 antagonistic interactions of different drugs with gentamicin or ciprofloxacin in *E. coli* to investigate the extent to which this occurs. Although initially detected at a growth inhibition level, all antagonisms held true at a killing level, with 14 of the 16 antagonisms working (at least partially) via decreasing the intracellular gentamicin or ciprofloxacin concentrations (Extended Data Fig. 8b). In several of the cases that we tested, this probably occurred because the second drug decreased the proton motive force-energized uptake of gentamicin or increased the AcrAB-TolC-dependent efflux of ciprofloxacin, as antagonisms were neutralized in the respective mutant backgrounds

within category. Edges represent synergy (blue) or antagonism (orange), and thickness reflects the number of interactions. Interactions between drugs of the same category or general cellular target are represented by self-interacting edges. Conserved interactions, including weak interactions, are shown. LPS, lipopolysaccharide. **d, e**, Count of synergistic and antagonistic drug–drug interactions in *E. coli*. Antagonisms occur almost exclusively between drugs that belong to different categories (**d**, across) or target different cellular processes (**e**, across), whereas synergies are also abundant between drugs within the same category (**d**) or that target the same process (**e**). χ^2 -test *P* values are shown for the difference in frequency of synergies over antagonisms between the ‘within’ and ‘across’ groups.

(Extended Data Fig. 8c). Overall, our results suggest that a large fraction of antagonisms is due to modulation of intracellular drug concentrations, rather than to direct interactions of the primary drug targets (Extended Data Fig. 8d, e).

Unlike antagonistic interactions, synergies often occurred between drugs that target the same cellular process (Fig. 1b–e, Extended Data Fig. 7). In fact, synergies are significantly enriched within drugs of the same category across all three species ($P < 10^{-16}$, Fisher’s exact test), given that in our dataset there are about 15-fold-more possible drug combinations across drug categories than within them. Mechanistically, by targeting the same functional process at different steps, drug combinations could bypass the redundancies of this process and thus have a synergistic effect. For example, the many synergies that exist between different β -lactams are probably because of their different affinities to the numerous and often-redundant penicillin-binding proteins (Fig. 1b, Extended Data Figs. 4e, 7a, b).

As with antagonisms, synergies can also occur owing to modulation of intracellular drug concentrations. Consistent with a general permeabilization role of membrane-targeting compounds in many organisms^{9,13,14} and with drug uptake being a major bottleneck for Gram-negative pathogens, one quarter of all the synergies that we detected contained at least one out of the eight membrane-targeting drugs used in our screen (two-sided Wilcoxon rank-sum test, $P = 0.06$). However, membrane-targeting compounds also account for about 18% of antagonisms, which suggests that perturbations in membrane integrity can also decrease intracellular drug concentrations. Consistently, benzalkonium decreases the intracellular concentration of both gentamicin and ciprofloxacin, probably by interfering with their import into the cell (Extended Data Fig. 8b, c).

We next examined the conservation of drug–drug interactions. Interactions within species were highly conserved (Fig. 2a, Extended Data Fig. 9a, b), with conservation being 53–76% depending on the species (Fig. 2b). Conservation is actually higher (68–87%, on average 80%) if we disregard non-comparable interactions for which the concentration range tested precluded us from detecting synergy or antagonism for both strains (Fig. 2b, Extended Data Fig. 3d). The high conservation of

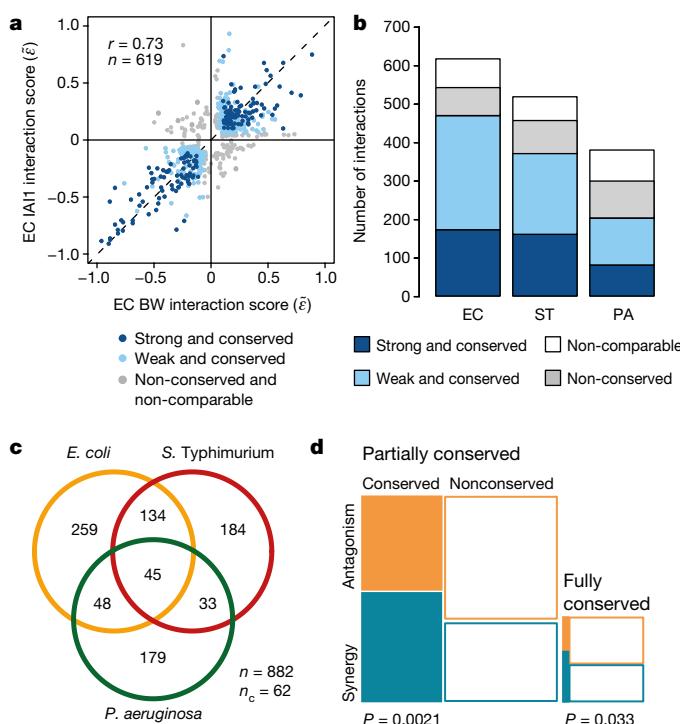


Fig. 2 | Drug-drug interaction conservation. **a**, Drug-drug interactions are conserved in *E. coli*. Scatter plot of interaction scores from the two *E. coli* strains; strong interactions for at least one of the strains are shown. Dark blue, strong and conserved interactions in both strains; light blue, strong interactions in one strain and concordant behaviour in other (weak and conserved); grey, interactions occurring exclusively in one strain or conflicting between strains (non-conserved or non-comparable, which refers to combinations that have considerably different single-drug dose responses between strains (Methods)). r denotes the Pearson correlation, n denotes the number of interactions plotted. **b**, Drug-drug interactions are highly conserved within species. Colours as in **a**; but non-conserved and non-comparable are separated. **c**, Drug-drug interactions are largely species-specific; n = total number of interactions; n_c = conflicting interactions between species, not accounted for in the Venn diagram. **d**, Synergies are more conserved than antagonisms. Mosaic plots and χ^2 -test P values correspond to the quantification of synergy and antagonism among conserved (fully and partially) and non-conserved interactions between species. EC, *E. coli*; EC BW, *E. coli* strain K-12 BW25113; EC IAI1, *E. coli* strain O8 IAI1; ST, *S. Typhimurium*; PA, *P. aeruginosa*.

drug-drug interactions within species is consistent with the finding that these interactions are generally robust to simple genetic perturbations¹⁵. Despite the high degree of conservation within species, 13–32% of the interactions remained strain-specific, with the majority being neutral in the second strain. Very few drug combinations synergized for one strain and antagonized for the other (16 interactions), but such strain differences held in our validation set (Supplementary Table 2).

Although conservation is relatively high within species, it is very low across species (Fig. 2c, Extended Data Fig. 9c). The majority (70%) of interactions occurred in only one species and only 5% were conserved in all three species, despite their close phylogenetic relationship. Because conservation is much higher at the single-drug level across the three species—which share resistance or sensitivity to 73% of the drugs (Supplementary Table 1, Methods)—this indicates that drug combinations can impart species specificity to the drug action. Such specificities can be beneficial for creating narrow-spectrum therapies with low collateral damage, by using synergies that are specific to pathogens and antagonisms that are specific to abundant commensals.

Moreover, we found that despite synergies being less prevalent (Fig. 1a), they are significantly more conserved than antagonisms (Fig. 2d). This is presumably because (i) synergies are enriched between drugs of the same category, and interactions within functional processes

are conserved across evolution¹⁶; (ii) membrane-targeting drugs have a general potentiation effect in Gram-negative bacteria; and (iii) antagonisms often depend on drug import or uptake (Extended Data Fig. 8), which are controlled by less-conserved envelope machineries.

Exploring the network of conserved drug–drug interactions across the three species (Extended Data Fig. 9d) exposed several potential Achilles' heels of Gram-negative bacteria, such as the strong synergy of colistin with macrolides¹⁷, and revealed that known antibiotic classes often behave non-uniformly. For example, the well-known synergy between β -lactams and aminoglycosides is confined to the potent aminoglycosides used in our screen (amikacin and tobramycin) and β -lactams (piperacillin, aztreonam and cefotaxime) that specifically target the cell-division-related penicillin-binding proteins, consistent with previous reports¹⁸. To address whether pairwise drug interactions are driven by mode of action (that is, drug classes interacting in a purely synergistic or antagonistic manner with one another)⁸, we calculated a monochromaticity index for all drug category pairs, across all species (Methods). For highly monochromatic category pairs, the monochromaticity index approaches 1 and -1 for antagonism and synergy, respectively. The monochromaticity index is high overall, especially between well-defined antibiotic classes. Yet, a number of these classes—including β -lactams, tetracyclines and macrolides—have mixed antagonisms and synergies with other antibiotic classes (Extended Data Fig. 9e). β -lactams have diverse affinities to their multiple penicillin-binding-protein targets (potentially explaining the mixed interactions with other classes) but the same does not apply to protein synthesis inhibitors, which have unique targets. In this case, non-uniform class behaviour may be due to different chemical properties of the class members, and thus different dependencies on uptake and efflux systems. Aggregating the monochromaticity index per drug category reinforced the view that broader categories exhibit less concordant interactions (Extended Data Fig. 9f). Besides membrane-targeting drugs, human-targeted drugs were the category that exhibited the most synergies, which suggests that many human-targeted drugs may act as adjuvants.

Because antibiotic classes interacted largely in a monochromatic fashion, clustering drugs according to their interactions recapitulated the class groupings (Extended Data Fig. 10). For example, cell-wall inhibitors grouped together, with further subdivisions being reflective of target specificity. However, exceptions—such as the macrolides, which split—were also evident. Azithromycin, the only dibasic macrolide, separates from its class co-members and clusters with two other basic antibiotics, bleomycin and phleomycin. Compared with other macrolides, azithromycin interacts with and crosses the outer membrane of Gram-negative bacteria in a distinct manner^{17,19} and also has different binding kinetics with the peptide exit tunnel of the 50S ribosomal subunit²⁰. For drugs with unknown or less-well-defined targets, clustering hinted at possible modes of action. Among them, we selected the flavouring compound vanillin, which clusters together with the structurally related acetyl-salicylic acid (aspirin). Salicylate and aspirin induce the expression of the major efflux pump in enterobacteria, AcrAB-TolC, by binding and inactivating the transcriptional repressor MarR²¹ (Fig. 3a). Consistent with a similar action, vanillin treatment increased levels of AcrA protein in *E. coli*, owing to *marA* overexpression (Fig. 3b, c). Higher AcrA levels upon vanillin or aspirin treatment led to higher minimal inhibitory concentrations for chloramphenicol and ciprofloxacin (Fig. 3d, e). As previously reported for salicylate²², vanillin exerts an additional minor effect on drug resistance in a MarR- and MarA-independent manner (Fig. 3c–e), presumably via the MarA homologue Rob.

To test whether detected interactions are relevant for resistant isolates, we selected seven strong and conserved synergies—comprising antibiotics, human-targeted drugs or food additives—and assessed their efficacy against six MDR *E. coli* and *Klebsiella pneumoniae* clinical isolates in total. All strains were recovered from patients with infections, and belong to successfully spread clonal lineages that contain extended spectrum β -lactamase resistance and various highly prevalent carbapenemases^{23,24}. One *K. pneumoniae* strain (929) is also resistant to the last-resort antibiotic, colistin, owing to a chromosomal mutation

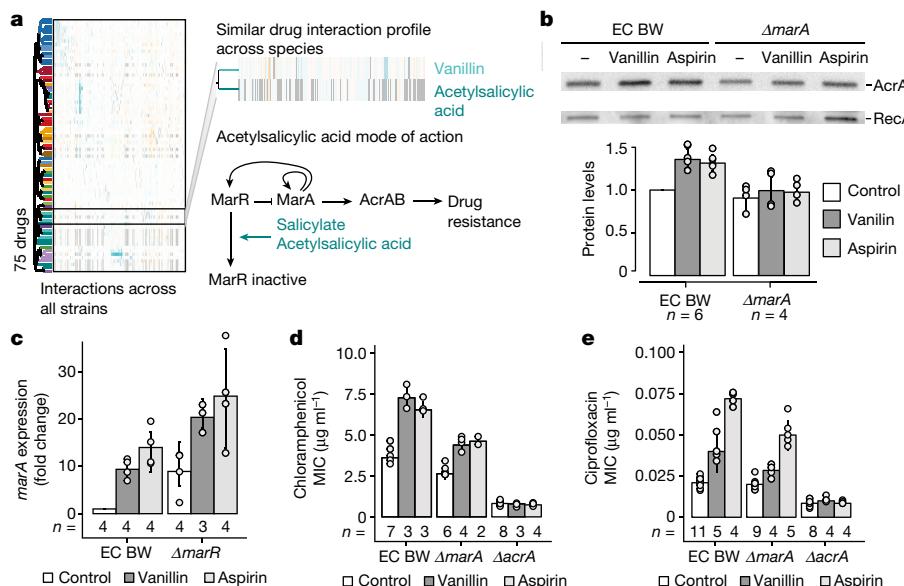


Fig. 3 | Vanillin induces a multi-antibiotic-resistance phenotype.

a, Vanillin and aspirin (acetylsalicylic acid) have similar drug–drug interaction profiles (see Extended Data Fig. 10), which suggests that they have similar modes of action. A schematic of the induction of the multi-antibiotic resistance response through deactivation of the MarR repressor by salicylate or aspirin²¹ is illustrated. b, Vanillin increases AcrA levels in a *marA*-dependent manner. A representative immunoblot of exponentially growing cells (all blots shown in Supplementary Fig. 1) after treatment with solvent, vanillin ($150 \mu\text{g ml}^{-1}$) or aspirin ($500 \mu\text{g ml}^{-1}$) is shown; loading was controlled by cell density and constitutively expressed RecA.

(Supplementary Table 4) that puts the strain in the category of extensively drug-resistant isolates. All drug pairs acted synergistically in most of the strains that we tested (Fig. 4a, Extended Data Fig. 11a). We further tested colistin–clarithromycin and spectinomycin–vanillin with an established infection model for evaluating antibacterial activity, using the larvae of *G. mellonella*. Both combinations also acted synergistically *in vivo* by increasing the rates of *G. mellonella* survival during infection (Fig. 4b, Extended Data Fig. 11b).

The strongest of these synergies is between colistin and different macrolides (Fig. 4, Extended Data Fig. 11). Although other polymyxins are known to help macrolides to cross the outer membrane of Gram-negative bacteria¹⁷, this particular synergy occurred at low colistin concentrations ($<0.3 \mu\text{g ml}^{-1}$) and was active even for the intrinsically colistin-resistant strain (*K. pneumoniae* 929, Fig. 4), which implies that macrolides may also potentiate colistin via an as-yet-unknown mechanism. Similar resensitization of colistin-resistant pathogens to colistin

by macrolides was recently reported for plasmid-borne colistin resistance²⁵, indicating that this synergy is independent of the resistance mechanism. In addition to antibiotic pairs, combinations of human-targeted drugs or food additives with antibiotics were also effective against MDR isolates, even when the former lacked antibacterial activity on their own (Extended Data Fig. 11).

One such compound, vanillin, potentiated the activity of spectinomycin in *E. coli* MDR isolates. This was intriguing, because vanillin antagonizes many other drugs including other aminoglycosides (Supplementary Table 2). We confirmed that this interaction is specific to spectinomycin and vanillin, and not to other aminoglycosides or aspirin, and thus independent of the vanillin effect on AcrAB–TolC (Extended Data Fig. 12a–c). We then probed a genome-wide *E. coli* gene knockout library²⁶ to identify mutations that abrogate the vanillin–spectinomycin interaction, but do not influence the interaction between vanillin and amikacin (another aminoglycoside).

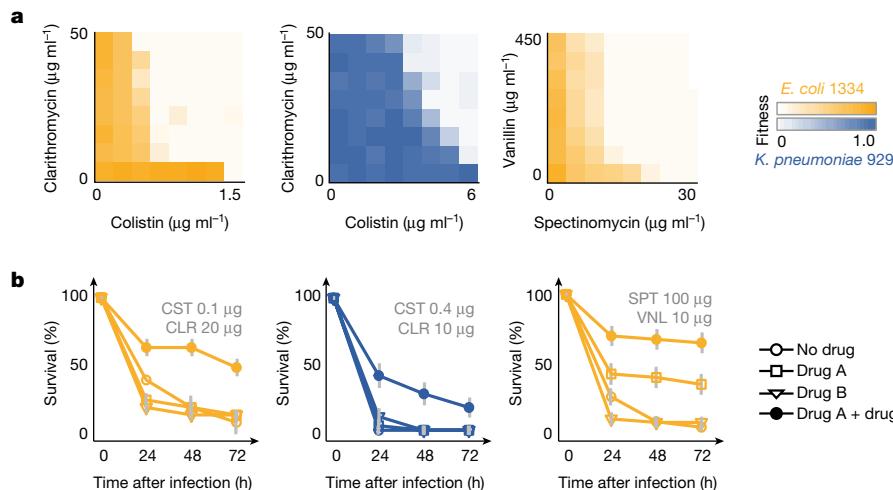


Fig. 4 | Potent synergistic combinations against Gram-negative MDR clinical isolates. a, In vitro synergies, shown as 8 × 8 checkerboards, for 2 MDR strains (more strains and synergies are shown in Extended Data Fig. 11). One of two biological replicates is shown. b, Drug synergies against the same MDR strains and drug combinations as in a in the *G. mellonella* infection model (see Extended Data Fig. 11). Larvae were infected by *E. coli* and *K. pneumoniae* MDR isolates (10^6 and 10^4 colony-forming units, respectively) and left untreated, treated with single drugs or with a combination of drugs. The percentage of surviving larvae was monitored at indicated intervals after infection. $n = 10$ larvae per treatment. The average of four biological replicates is shown; error bars depict s.d. CST, colistin; CLR, clarithromycin; SPT, spectinomycin; VNL, vanillin.

One of the top hits was *mdfA*, which encodes a transporter of the major facilitator superfamily that exports both charged and neutral compounds²⁷ (Extended Data Fig. 12c). Consistent with MdfA modulating spectinomycin uptake, $\Delta mdfA$ (*mdfA* deletion mutant) cells were more resistant to spectinomycin and not responsive to vanillin (Extended Data Fig. 12d), whereas cells overexpressing *mdfA* were more sensitive to spectinomycin (Extended Data Fig. 12d, e), as previously reported²⁸, with vanillin further exacerbating this effect (Extended Data Fig. 12d). Vanillin addition also increased the intracellular spectinomycin concentration in an *mdfA*-dependent manner (Extended Data Fig. 12f). At this point, it is unclear how MdfA—which is known to export compounds out of the cell—facilitates spectinomycin import into the cell. However, the phylogenetic occurrence of *mdfA* is concordant with the species-specificity of this interaction, as we detected this synergy in *E. coli* and *S. Typhimurium* but not in *P. aeruginosa* and *K. pneumoniae* isolates, which lack *mdfA*. This synergy underlines the importance of exploring the role of food additives in combinatorial therapies⁵.

In summary, we generated a comprehensive resource of pairwise drug combinations in Gram-negative bacteria, which illuminates key principles of drug–drug interactions and provides a framework for assessing their conservation across organisms or individuals (Supplementary Discussion). Such information can form the basis for similar screens in other microbes, studies investigating the underlying mechanism of pairwise drug combinations^{11,15,29} and computational predictions of their outcomes^{30,31}. Some of the principles that we have identified probably go beyond anti-infectives and microbes³². For anti-bacterial drug therapies, our study shows that non-antibiotic drugs hold promise as adjuvants, offers a new path for narrow spectrum therapies and identifies effective synergies against MDR clinical isolates (Supplementary Discussion). Further experimentation is required to address whether such synergies have clinical relevance.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0278-9>.

Received: 11 May 2017; Accepted: 24 May 2018;

Published online 4 July 2018.

- President of the General Assembly of the United Nations. Press release: high-level meeting on antimicrobial resistance (<http://www.un.org/pga/71/2016/09/21/press-release-hl-meeting-on-antimicrobial-resistance/>) (2016).
- Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
- Tacconelli, E. et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327 (2018).
- Ejim, L. et al. Combinations of antibiotics and nonantibiotic drugs enhance antimicrobial efficacy. *Nat. Chem. Biol.* **7**, 348–350 (2011).
- Brown, D. Antibiotic resistance breakers: can repurposed drugs fill the antibiotic discovery void? *Nat. Rev. Drug Discov.* **14**, 821–832 (2015).
- Kohanski, M. A., Dwyer, D. J., Hayete, B., Lawrence, C. A. & Collins, J. J. A common mechanism of cellular death induced by bactericidal antibiotics. *Cell* **130**, 797–810 (2007).
- Erzaty, B. et al. Fe–S cluster biosynthesis controls uptake of aminoglycosides in a ROS-less death pathway. *Science* **340**, 1583–1587 (2013).
- Yeh, P., Tschumi, A. I. & Kishony, R. Functional classification of drugs by properties of their pairwise interactions. *Nat. Genet.* **38**, 489–494 (2006).
- Robbins, N. et al. An antifungal combination matrix identifies a rich pool of adjuvant molecules that enhance drug activity against diverse fungal pathogens. *Cell Reports* **13**, 1481–1492 (2015).
- Cokol, M. et al. Large-scale identification and analysis of suppressive drug interactions. *Chem. Biol.* **21**, 541–551 (2014).
- Bollenbach, T., Quan, S., Chait, R. & Kishony, R. Nonoptimal microbial response to antibiotics underlies suppressive drug interactions. *Cell* **139**, 707–718 (2009).
- Costanzo, M. et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).
- Farha, M. A. & Brown, E. D. Chemical probes of *Escherichia coli* uncovered through chemical–chemical interaction profiling with compounds of known biological activity. *Chem. Biol.* **17**, 852–862 (2010).
- Stokes, J. M. et al. Pentamidine sensitizes Gram-negative pathogens to antibiotics and overcomes acquired colistin resistance. *Nat. Microbiol.* **2**, 17028 (2017).

- Chevereau, G. & Bollenbach, T. Systematic discovery of drug interaction mechanisms. *Mol. Syst. Biol.* **11**, 807 (2015).
- Ryan, C. J. et al. Hierarchical modularity and the evolution of genetic interactomes across species. *Mol. Cell* **46**, 691–704 (2012).
- Vaara, M. Outer membrane permeability barrier to azithromycin, clarithromycin, and roxithromycin in Gram-negative enteric bacteria. *Antimicrob. Agents Chemother.* **37**, 354–356 (1993).
- Giannouli, H., Zissis, N. P., Tagari, G. & Bouzos, J. In vitro synergistic activities of aminoglycosides and new beta-lactams against multiresistant *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **25**, 534–536 (1984).
- Imamura, Y. et al. Azithromycin exhibits bactericidal effects on *Pseudomonas aeruginosa* through interaction with the outer membrane. *Antimicrob. Agents Chemother.* **49**, 1377–1380 (2005).
- Petropoulos, A. D. et al. Time-resolved binding of azithromycin to *Escherichia coli* ribosomes. *J. Mol. Biol.* **385**, 1179–1192 (2009).
- Hao, Z. et al. The multiple antibiotic resistance regulator MarR is a copper sensor in *Escherichia coli*. *Nat. Chem. Biol.* **10**, 21–28 (2014).
- Chubiz, L. M., Glekas, G. D. & Rao, C. V. Transcriptional cross talk within the *mar-sox-rob* regulon in *Escherichia coli* is limited to the *rob* and *marRAB* operons. *J. Bacteriol.* **194**, 4867–4875 (2012).
- Göttig, S., Hamprecht, A. G., Christ, S., Kempf, V. A. & Wichelhaus, T. A. Detection of NDM-7 in Germany, a new variant of the New Delhi metallo-β-lactamase with increased carbapenemase activity. *J. Antimicrob. Chemother.* **68**, 1737–1740 (2013).
- Göttig, S., Gruber, T. M., Stecher, B., Wichelhaus, T. A. & Kempf, V. A. In vivo horizontal gene transfer of the carbapenemase OXA-48 during a nosocomial outbreak. *Clin. Infect. Dis.* **60**, 1808–1815 (2015).
- MacNair, C. R. et al. Overcoming *mcr-1* mediated colistin resistance with colistin in combination with other antibiotics. *Nat. Commun.* **9**, 458 (2018).
- Baba, T. et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
- Yardeni, E. H., Zomot, E. & Bibi, E. The fascinating but mysterious mechanistic aspects of multidrug transport by MdfA from *Escherichia coli*. *Res. Microbiol.* <https://doi.org/10.1016/j.resmic.2017.09.004> (2017).
- Bohn, C. & Bouloc, P. The *Escherichia coli* *cmlA* gene encodes the multidrug efflux pump Cmr/MdfA and is responsible for isopropyl-β-d-thiogalactopyranoside exclusion and spectinomycin sensitivity. *J. Bacteriol.* **180**, 6072–6075 (1998).
- Nichols, R. J. et al. Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
- Wildenhain, J. et al. Prediction of synergism from chemical–genetic interactions by machine learning. *Cell Syst.* **1**, 383–395 (2015).
- Chandrasekaran, S. et al. Chemogenomics and orthology-based design of antibiotic combination therapies. *Mol. Syst. Biol.* **12**, 872 (2016).
- Lehár, J. et al. Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.* **27**, 659–666 (2009).

Acknowledgements We thank P. Beltrao (EBI) and T. Bollenbach (University of Cologne) for providing feedback on the manuscript; K. M. Pos (Goethe University) for the anti-AcrA antibody; D. Helm and the EMBL Proteomics Core Facility for help with mass spectrometry experiments; the EMBL GBCS and the Centre for Statistical Analysis for advice on data analysis; S. Riedel-Christ for help with *G. mellonella* experiments; and the members of the Typas laboratory for discussions. This work was partially supported by EMBL internal funding, the Sofja Kovalevskaja Award of the Alexander von Humboldt Foundation to A.Ty., the JPIAMR Combinatorials grant to F.B. (ANR) and A.Ty. (BMBF), and the DFG (FOR 2251) to S.G. A.M. and J.S. are supported by a fellowship from the EMBL Interdisciplinary Postdoc (EIPOD) program under Marie Curie Actions COFUND.

Reviewer information *Nature* thanks A. Koul, K. Lewis and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.R.B. and A.Ty. conceived and designed the study. A.R.B., A.Ty. and J.B. performed the screen; A.R.B., A.Ty. and N.N. the validation screen; and A.R.B., M.B., A.M., J.S., S.B., M.Z. and J.Z.B. the mechanistic follow-up work. S.G. characterized the clinical isolates. A.R.B., A.Ty. and S.F. performed the clinical isolate checkerboards, and E.H. and S.G. performed the *G. mellonella* infection experiments. A.R.B. analysed all data. B.P., F.B., S.G. and A.Ty. supervised different parts of this study; B.E., M.M.S. and P.B. provided advice. A.R.B. and A.Ty. wrote the paper with input from M.M.S., P.B. and S.G. All authors approved the final version.

Competing interests EMBL has filed a patent application on using drug combinations identified in this study for prevention and/or treatment of infections and antibacterial-induced dysfunctions (European patent application number EP18169989.3). A.B., S.G. and A.Ty. are listed as inventors.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0278-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0278-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.Ty.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size.

Strains, plasmids and drugs. For each of the three Gram-negative species profiled in this study, we used two common sequenced laboratory strains: *E. coli* K-12 BW25113 and O8 IAI1 (hereafter IAI1), *S. Typhimurium* LT2 and 14028s, and *P. aeruginosa* PAO1 and PA14. To validate selected synergies, we profiled 6 MDR clinical Enterobacteriaceae isolates recovered from specimens from human patients: *E. coli* 124, 1027 and 1334, and *K. pneumoniae* 718, 929 and 980 (see Supplementary Table 4 for details of antibiotic resistance determinants). For follow-up experiments, we used two closely related *E. coli* K-12 model strains, BW25113 and MG1655.

All mutants used in this study were made using the *E. coli* Keio knockout collection²⁶, after confirming with PCR and retransducing the mutation to wild-type BW25113 with the P1 phage (Supplementary Table 5). The kanamycin resistance cassette was excised when necessary using the plasmid pCP20³³. The plasmid used for *mdfA* overexpression was obtained from the mobile *E. coli* ORF library³⁴.

Drugs used in this study were purchased from Sigma Aldrich, except for metformin hydrochloride (TCI Chemicals), clindamycin and bleomycin (Applichem), CHIR-090 (MedChemtronic) and vanillin (Roth). Stocks were prepared according to the supplier's recommendations (preferably dissolved in water).

Minimal inhibitory concentration calculation. We defined minimal inhibitory concentration (MIC) as the lowest concentration required to inhibit growth of a microorganism after 8 h of incubation in lysogeny broth (LB) at 37 °C with shaking (384-well plates, starting with an optical density at 595 nm (OD_{595 nm}) of 0.01). MICs of all drugs were computed using a logistic fit of growth (OD_{595 nm} for 8 h) over twofold serial dilutions of the antibiotic concentrations for all strains used for the high-throughput screening and follow-up experiments.

High-throughput screening of pairwise drug interactions. For all drug combination experiments, drugs were diluted in LB to the appropriate working concentrations in transparent 384-well plates (Greiner BioOne GmbH), with each well containing 30 µl in total. After the addition of drugs, cells were inoculated at initial OD_{595 nm} of about 0.01 from an overnight culture. The same inoculum was used for all strains. All liquid handling (drug addition and cell mixing) was done with a Biomek FX liquid handler (Beckman Coulter). Plates were sealed with breathable membranes (Breathe-Easy) and incubated at 37 °C in a humidity-saturated incubator (Cytomat 2, Thermo Scientific) with continuous shaking and without lids to avoid condensation. OD_{595 nm} was measured every 40 min for 12 h in a Filtermax F5 multimode plate reader (Molecular Devices).

A flowchart of the experimental and analytical pipeline is shown in Extended Data Fig. 2a. Data analysis was implemented with R and networks were created with Cytoscape³⁵.

Experimental pipeline. The drug–drug interaction screen was performed using 4 × 4 checkerboards. Sixty-two drugs were arrayed in 384-well plates with the different concentrations in duplicates (array drugs). Each plate contained 12 randomly distributed wells without arrayed drug: 9 wells containing only the query drug, and 3 wells without any drug. One query drug at a single concentration was added to all wells of the 384-well plate, except for the 3 control wells. All drugs were queried once—or occasionally twice—per concentration. We used 78 drugs as query in *E. coli* and *S. Typhimurium*, and 76 in *P. aeruginosa*. In total, 79 query drugs were screened, out of which 75 were common to all three species (Supplementary Table 1). The 62 array drugs were a subset of the 79 query drugs. The same drug concentrations were used in both query and array drugs (Supplementary Table 1). Three drug concentrations (twofold dilution series) were selected based on the MIC curves, tailored to the strain and drug. We targeted for nearly full, moderate and mild or no growth inhibition, which on average corresponded to 50–100%, 25–50% and 0–25% of the MIC, respectively. The highest drug concentration and the lowest fitness obtained per single drug are listed in Supplementary Table 1. For drugs that do not inhibit growth on their own, we selected concentrations according to sensitivity of other strains or species, or to their use in clinics or for research. *E. coli* and *S. Typhimurium* exhibited largely similar single-drug dose responses within species, thus the same drug concentrations were used for both strains of each species. For *P. aeruginosa*, MICs often differed by several fold and drug concentrations were therefore adjusted between the two strains (Supplementary Table 1).

Growth curve analysis. The Gompertz model was fitted to all growth curves (when growth was observed) by using the R package grofit version 1.1.1-1 for noise reduction. Quality of fit was assessed by Pearson correlation (*r*), which was >0.95 for approximately 95% of all growth curves. *r* < 0.95 was indicative of either non-sigmoidal-shaped growth curves—typical of some drugs such as fosfomycin—or noisy data. Noisy data were removed from further analysis. Plate effects were corrected by fitting a polynomial to the median growth of each row and column. The background signal from LB was removed by subtracting the median curve of the non-growing wells from the same plate. These were wells in which either the single- or the double-drug treatments fully inhibited growth; each plate contained

at least three such wells. Data were processed per strain and per batch to correct for systematic effects.

Fitness estimation. We used a single time-point OD_{595 nm} measurement (growth) for assessing fitness. This corresponded to the transition to stationary phase for cells grown without perturbation, as this enables us to capture the effect of drugs on lag-phase, growth rate or maximum growth. Thus, we used OD_{595 nm} at 8 h for *E. coli* BW25113 and both *P. aeruginosa* strains, at 7 h for the fast-growing *E. coli* IAI1 and *S. Typhimurium* 14028s, and at 9 h for the slower-growing *S. Typhimurium* LT2.

We used the Bliss model to assess interactions as it can accommodate drugs that have no effect alone, but which potentiate the activity of others (adjutants)³⁶. This feature is especially relevant here, because we probed intrinsically antibiotic-resistant microbes (*P. aeruginosa* and MDR clinical isolates), and human-targeted drugs or food additives that lack antibacterial activity. According to the Bliss independence model³⁷ and assuming that drug–drug interactions are rare, for most drug combinations the fitness of arrayed drugs (*f_a*) equals the fitness in the presence of both drugs (*f_{aq}*) divided by the fitness of the query drug alone (*f_q*):

$$\varepsilon = f_{\text{aq}} - f_{\text{a}} \times f_{\text{q}} \quad (1)$$

If $\varepsilon = 0$,

$$f_{\text{a}} = \frac{f_{\text{aq}}}{f_{\text{q}}} \Leftrightarrow f_{\text{a}} = \frac{g_{\text{aq}}/g_0}{g_{\text{q}}/g_0} \Leftrightarrow f_{\text{a}} = \frac{g_{\text{aq}}}{g_{\text{q}}} \quad (2)$$

in which ε denotes the Bliss score, f denotes fitness, g denotes growth, subscript a denotes that the variable pertains to an arrayed drug, subscript q denotes that the variable pertains to a query drug and 0 denotes no drug. The fitness in the presence of both drugs (*f_{aq}*) was calculated by dividing the growth in the presence of both drugs (*g_{aq}*) by the median of the growth of drug-free wells from the same plate (*g₀*). The fitness of the single query drugs (*f_q*) was obtained by dividing the top 5% growing wells across each batch by the median of the growth of drug-free wells of each plate (*g₀*). This metric is more robust to experimental errors than using only the 9 wells containing the query drug alone. Nevertheless, both estimators for *f_q* yield very similar results (Pearson correlation, *r* = 0.98). Consistent with equation (2), the fitness of arrayed drugs (*f_a*) was estimated by the slope of the line of best fit between *g_{aq}* and *g_q* across all plates (query drugs) within a batch.

$$\begin{bmatrix} g_{\text{q}_1} \\ \vdots \\ g_{\text{q}_n} \end{bmatrix}_{n \times 1} \times f_{\text{a}_m} = \begin{bmatrix} g_{\text{a}_m \text{q}_1} \\ \vdots \\ g_{\text{a}_m \text{q}_n} \end{bmatrix}_{n \times 1} \quad 1 \leq m \leq \text{nr} \text{ (number of arrayed drugs)}$$

for a given array drug *m* (*a_m*) across *n* query drugs (*q_n*) within a batch (Extended Data Fig. 2b).

For array drugs with a Pearson correlation (*r*) between *g_{aq}* and *g_q* below 0.7, *f_a* was estimated using only the query drugs that corresponded to the interquartile range of *g_{aq}/g_q* (minimum *n* = 18 query drugs, Extended Data Fig. 2b). Wells for which *r* was still below 0.7, even after restricting the number of plates were removed from further analysis owing to high noise (~2%). For wells exhibiting no growth for >75% of the plates within a batch, *f_a* was deemed to be zero.

Bliss independence interaction scores. Bliss scores (ε) were calculated for each well, as described in equation (1). At least 3 × 3 drug concentrations × 2 (duplicates) × 2 (query and array drugs) = 36, or 18 (drugs used only as query) scores were obtained per drug pair. Drug–drug interactions were inferred based on the Bliss independence model in three steps: (i) strong interactions based on complete ε distributions, (ii) strong interactions based on ε distributions restricted to relevant drug concentrations and (iii) weak and conserved interactions within species. Cross-species comparisons, drug–drug interaction networks and monochromaticity analyses shown in this study include all drug–drug interactions.

Strong drug–drug interactions based on complete ε distributions. Strong drug–drug interactions were statistically assigned using a re-sampling approach. Ten thousand repetitions of a two-sided Wilcoxon rank-sum test (per drug pair and per strain) were performed, to sample a representative set of ε for a given strain. For every repetition, the ε distribution of a given combination was compared to an ε distribution of the same size randomly sampled from the complete ε set for a given strain. *P* values were calculated as follows:

$$P = \frac{\sum_{n=1}^N (P_n > 0.1) + 1}{N + 1}$$

in which *N* is the total number of repetitions (10,000) and *P_n* is the *P* value of the Wilcoxon rank-sum test obtained for the *n*th repetition. Strong drug–drug interactions were assigned to drug pairs that simultaneously satisfied two criteria: (i) first or third quartile of the ε distribution below −0.1 or higher than 0.1 for synergies or antagonisms, respectively, and (ii) *P* < 0.05 (after correcting for multiple

testing, Benjamini–Hochberg). Only one-sided drug interactions were taken into account, thus the very few interactions that satisfied the criteria concurrently for synergy and antagonism were re-assigned as neutral (only $n = 1$ for $|\tilde{\varepsilon}| > 0.1$). The highest absolute ε value between the first and third quartiles was used as single interaction score ($\tilde{\varepsilon}$) to reflect the strength of the drug–drug interactions.

Strong drug–drug interactions based on ε distributions restricted to relevant drug concentrations. Because drug interactions are dependent on concentration, the same statistical procedure was repeated after restricting the drug concentration ratios to those relevant for either synergy or antagonism. This constraint was added by excluding ε values that corresponded to concentration ratios in which the expected fitness (product of the fitness on single drugs, $f_x \times f_y$) was below 0.2 for synergy and above 0.8 for antagonism, which represent blind spots for the given interaction type (Extended Data Fig. 3d). These interactions are described by their P value and $\tilde{\varepsilon}$ obtained with restricted drug concentration ratios. Although most interactions were detected based on both full and restricted ε distributions, each of the different methods uniquely identified interactions (Extended Data Fig. 4c). With the expected fitness cutoff of 0.2, we identified the highest number of strong interactions (1,950) with 90 uniquely identified interactions from full ε distributions and 379 from restricted ε distributions (see also ‘Sensitivity analysis’).

Restricting ε values based on expected fitness also enables defining whether synergy or antagonism is detectable for any given drug pair. No significant P value was found for drug pairs with less than five ε scores within the relevant expected fitness space, as their sample size is insufficient. Synergy and antagonism could not be detected for 1% and 25% of all drug combinations, respectively.

Weak and conserved drug–drug interactions within species. For drug pairs with a strong drug–drug interaction in only one of the two strains per species, the criteria for assigning interactions for the second strain was relaxed to $|\tilde{\varepsilon}_{\text{second strain}}| > 0.06$, provided that the interaction sign was the same. Interactions assigned with this approach are termed weak and conserved.

Loewe additivity interaction scores. For combinations between β -lactams for which high-resolution 8×8 checkerboards with sufficient growth inhibition were available in the validation dataset, Loewe additivity³⁸ was used to confirm the interactions. Drug–drug interactions were inferred by the shape of the isoboles (lines of equal growth) in two-dimensional drug-concentration plots. Unless stated otherwise, all isoboles correspond to 50% growth inhibition (IC_{50}) and were obtained by fitting a logistic model, with lines representing isoboles and dots representing IC_{50} interpolated concentrations. To interpolate IC_{50} concentrations (or other percentages of growth inhibition), a logistic model was used to fit the growth for each concentration of the first drug across different concentrations of the second drug. The null hypothesis of this model is represented by the additivity line: a linear isobole connecting equal individual growth inhibition values for the two drugs.

Sensitivity analysis. We confirmed the adequacy of the main statistical parameters used to assign interactions by performing a sensitivity analysis. Several expected fitness ($f_x \times f_y$) cutoffs were tested, while keeping the other parameters constant (Extended Data Fig. 4c). The added value of restricting the ε distributions to relevant drug concentrations (based on expected fitness) was strongly supported by the proportion of strong drug–drug interactions that was found exclusively using this criterion (~19% with our selected cutoff). The selected cutoff (0.2; disregarding wells with $f_x \times f_y < 0.2$ for synergies and with $f_x \times f_y > 0.8$ for antagonisms) resulted in the largest number of total interactions assigned, and the highest precision (91%) and recall (74%) after benchmarking against the validation dataset (Extended Data Fig. 4c).

The suitability of the thresholds applied to define strong ($|\tilde{\varepsilon}| > 0.1$) and weak ($|\tilde{\varepsilon}| > 0.06$) interactions was assessed by their effect on the true- and false-positive rates (Extended Data Fig. 4d). A threshold of $|\tilde{\varepsilon}| > 0.1$ is beneficial, as it imposes a minimum strength to assign interactions. A value of 0.1 corresponds to ~3 times the median of the first and third quartiles across all ε distributions (Extended Data Fig. 2c). Lowering this threshold results in lower true-positive rate, because several drug pairs are reassigned as neutral owing to ambiguity in calling interactions (we do not allow interactions to be both a synergy and an antagonism). Increasing this threshold lowers the true-positive rate, because only very strong interactions will be assigned (Extended Data Fig. 4d). Drug–drug interactions are highly conserved within species, which is evident from the high correlation for $\tilde{\varepsilon}$ that is observed for all species (Fig. 2a, Extended Data Fig. 9a, b). This motivated us to relax the interaction-strength threshold for the second strain if the interaction score $|\tilde{\varepsilon}|$ was above 0.1 in the first strain, dubbing these interactions weak and conserved. By including weak and conserved interactions in our analysis, the true-positive rate was increased by 15%. Adding a threshold for weak interactions of $|\tilde{\varepsilon}| > 0.06$ (about two times the median of the first and third quartiles of all ε distributions) is key for maintaining a low false-positive rate (Extended Data Fig. 4d).

Benchmarking and clinical isolates checkerboard assays. Eight-by-eight checkerboard assays were performed to validate our screen (242 drug combinations in the benchmarking dataset, Supplementary Table 3), and to test 7 selected synergies against 6 MDR clinical isolates (Fig. 4, Extended Data Fig. 11). As in the screen,

growth was assessed on the basis of $OD_{595\text{ nm}}$ at the transition to stationary phase for the no-drug controls. The time points used in the screen were used again for the validation set, and 8 h was used for all *E. coli* and *K. pneumoniae* MDR isolates. Fitness was calculated by dividing $OD_{595\text{ nm}}$ after single- or double-drug treatment by no-drug treatment for each individual checkerboard. Bliss scores (ε) were calculated as before, resulting in 49 ε values per drug pair. Drug combinations were analysed on the basis of ε distributions, after removing wells in which one of the drugs alone—and its subsequent combinations with the second drug—completely inhibited growth. Antagonism was assigned when the median of the ε distribution was above 0.1, or the third quartile was above 0.15. Similarly, synergies were assigned when the median of the ε distribution was below -0.1 or the first quartile was below -0.15. All experiments were done in biological duplicates, and interactions were considered effective when duplicates were consistent (as was the case for the vast majority of interactions).

Assessing conservation of drug–drug interactions. Conservation of drug–drug interactions between strains of the same species was assessed by Pearson correlation of the interactions scores, $\tilde{\varepsilon}$. For potentially non-conserved drug–drug interactions, the expected fitness distributions of the two strains were taken into account. When the two distributions were significantly different according to a two-sided Wilcoxon rank-sum test ($P < 0.05$ after Benjamini–Hochberg correction for multiple testing), the drug pairs were deemed as non-comparable between the two strains.

To assess the cross-species conservation of drug–drug interactions, we took into account only drug pairs that were probed in all three species. Drug–drug interactions were defined as being detected within a species when detected in at least one of the two strains and when no change of interaction sign was observed for the other strain. Interactions were then compared across the three species. Cases in which an interaction between drugs changed from synergy to antagonism or vice versa across species (conflicting interactions; ~7% of all interactions, Supplementary Table 2) were excluded from the comparative ‘across-species’ Venn diagram (Fig. 2c). In current analysis, a given drug–drug interaction may be conserved across species but not conserved within the species.

Conservation at the single-drug level was defined on the basis of shared resistance and sensitivity (Supplementary Table 1). A strain was considered sensitive to a given drug if one of the drug concentrations resulted in at least 30% growth inhibition. Consistent with conservation of drug–drug interactions across species, single-drug responses are conserved across species when at least one strain of both species has the same sign (sensitive or resistant).

Monochromaticity index. The monochromaticity index (MI) between drug pairs has previously been defined³⁹ as: if $r_{ij} > b$, then

$$MI_{ij} = \frac{(r_{ij} - b)}{1 - b}$$

if $r_{ij} = b$, then $MI_{ij} = 0$
and if $r_{ij} < b$, then

$$MI_{ij} = \frac{(r_{ij} - b)}{b}$$

in which r_{ij} denotes the ratio of antagonism to all interactions between drugs from classes i and j , and b denotes the ratio of antagonism to all interactions. We set a minimum of two interactions between drugs from classes i and j to calculate the monochromaticity index. The monochromaticity index equals 1 if only antagonisms occur between drugs from classes i and j , and -1 if only synergies occur between the classes. The monochromaticity index equals zero if the fraction of antagonism between the two classes reflects the background ratio b . Both strong and weak drug interactions were taken into account across all species, to obtain one monochromaticity index per drug category pair.

Assessment of drug combinations in the *G. mellonella* infection model. Larvae of the greater wax moth (*G. mellonella*) at their final instar larval stage were used as an *in vivo* model to assess efficacy of drug combinations. Larvae were purchased from UK Waxworms and TZ-Terraristik. Stock solutions of vanillin (in 20% DMSO), spectinomycin, colistin and clarithromycin (in 20% DMSO and 0.01% glacial acetic acid) were freshly prepared and diluted in PBS to the required concentration. Drugs and bacterial suspensions were administered by injection of 10- μ l aliquots into the haemocoel through the final pair of prolegs (bacteria into the left proleg, and antibiotics into the right), using Hamilton precision syringes. Controls included both uninfected larvae, and infected and uninfected larvae that were injected with the solvent used for the drugs. Drug toxicity was pre-evaluated by injection of serial dilutions of either single drugs or drug combinations, and drugs were used at amounts that caused little or no toxicity. To identify an optimal inoculum, time-kill curves were generated by inoculating larvae with 10 μ l of serially diluted bacterial suspensions (1×10^2 – 1×10^7 CFUs). For final experiments, groups of ten larvae were injected per strain-drug combination, placed into

Petri dishes and incubated at 37 °C. Larvae were infected with a sublethal dose of 1×10^6 and 1×10^4 CFUs for *E. coli* and *K. pneumoniae* isolates, respectively, and subsequently injected with indicated drugs, 1-h after infection. Survival of the larvae was monitored at the indicated time points by two observers independently. Each strain–drug combination was evaluated in four independent experiments.

Cell viability assays and intracellular antibiotic concentration. *Ciprofloxacin.* Overnight cultures of *E. coli* BW25113 were diluted 1:1,000 into 50 ml LB and grown at 37 °C to OD_{595 nm} ~ 0.5. Paraquat (50 µg/ml), vanillin (150 µg/ml), benzalkonium (5 µg/ml), caffeine (200 µg/ml), doxycycline (0.5 µg/ml), rifampicin (5 µg/ml), trimethoprim (5 µg/ml) or curcumin (100 µg/ml) were added to the cultures and incubated at 37 °C for 30 min. Ciprofloxacin (2.5 µg/ml final concentration) was then added to the cultures and cultures were incubated at 37 °C for 1 h in the presence of both drugs. Cell viability was determined by counting CFUs after plating washed cell pellets onto LB agar Petri dishes and incubating for 16 h. Intracellular ciprofloxacin was quantified using liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS), as previously described^{40,41}. Non-washed cell pellets⁴² were directly frozen and lysed with 350 µl of acetonitrile, followed by three freeze–thaw cycles (thawing was performed in an ultrasonic bath for 5 min). Cell debris was pelleted at 16,000g and the supernatant was filtered through a 0.22-µm syringe filter before injection. Chromatographic separation was performed on a Waters BEH C18 column (2.1 × 50 mm; 1.7 µm) at 40 °C, with a 2-min gradient with flow rate of 0.5 ml/min: (i) 0–0.5 min, 1% mobile phase B; (ii) 0.5–1.2 min, linear gradient from 1 to 95% mobile phase B; (iii) 1.2–1.6 min, 95% mobile phase B; and (iv) 1.6–1.7 min, return to initial conditions. Mobile phase A consisted of 0.1% formic acid in water, and mobile phase B consisted of 0.1% formic acid in acetonitrile. Samples were kept at 4 °C until analysis. Sample injection volume was 5 µl. Detection of ciprofloxacin was performed on a Waters Q–ToF premier instrument with electrospray ionization in positive mode. The transition 332 > 314 was monitored, with cone voltage set at 8 and collision energy set at 20. Intracellular ciprofloxacin was normalized to CFUs at the time of ciprofloxacin addition.

Gentamicin. Intracellular gentamicin was quantified by measuring [³H]-gentamicin (1 mCi/ml; Hartmann Analytic), as previously described⁷. Overnight cultures of *E. coli* MG1655 (the parental strain of BW25113) were diluted 1:100 into 5 ml LB and grown to OD_{595 nm} ~ 0.1. [³H]-Gentamicin was diluted in cold gentamicin to obtain a 5 µg/ml (0.1 mCi/ml) stock solution, which was then added to the culture at a final concentration of 5 µg/ml (0.1 µCi/ml) together with the second drug: berberine (200 µg/ml), erythromycin (15 µg/ml), metformin (13,000 µg/ml), procaine (6,000 µg/ml), loperamide (400 µg/ml), benzalkonium (5 µg/ml), rifampicin (5 µg/ml) or clindamycin (200 µg/ml). Cultures were then incubated at 37 °C on a rotary shaker. At 0, 0.5, 1, 1.5 and 2-h time-points, 500-µl aliquots were removed and applied to a 0.45-µm-pore-size HAWP membrane filter (Millipore) pretreated with 1 ml of unlabelled gentamicin (250 µg/ml). Filters were washed with 10 ml of 1.5% NaCl, placed into counting vials and dried for 30 min at 52 °C. Subsequently, 8 ml of liquid scintillation was then added to the dried filters and vials were incubated overnight at room temperature before being counted for 5 min. Gentamicin uptake efficiency is expressed as total accumulation of gentamicin (in ng) per 10^8 cells. Cell viability was determined by CFUs.

Spectinomycin. Intracellular spectinomycin was quantified by measuring [³H]-spectinomycin (1 µCi/mg; Hartmann Analytic). Overnight cultures of *E. coli* BW25113 were diluted 1:1,000 into 1 ml LB with and without vanillin (150 µg/ml) and grown to OD_{595 nm} ~ 0.5. Then, 50 µg/ml [³H]-spectinomycin:spectinomycin 1:100 was added and the cultures were incubated for 1 h. Cultures were pelleted, washed twice with PBS with 50 µg/ml non-labelled spectinomycin, re-suspended in 1% SDS and incubated for 20 min at 85 °C. The lysate was mixed with 8 ml liquid scintillation (Perkin Elmer ULTIMA Gold) and counted for 1 min using a Perkin Elmer Tri-Carb 2800TR. Measured radioactivity was normalized to cell number as measured by OD_{595 nm}.

RNA isolation, cDNA preparation and quantitative RT–PCR. Overnight cultures of *E. coli* BW25113 and the *marR* deletion mutant ($\Delta marR$) were diluted 1:2,000 into 20 ml LB and grown at 37 °C to OD_{595 nm} ~ 0.2. Aspirin or vanillin were added to the cultures to a final concentration of 500 and 150 µg/ml, respectively (DMSO was added in the control), followed by a 30-min incubation period at 37 °C with agitation. Cells were collected and RNA was extracted using the RNeasy Protect Bacteria Mini Kit (Qiagen). cDNA was prepared for RT–qPCR using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific). Levels of *marA* expression were estimated by RT–qPCR using SYBR Green PCR master mix following the manufacturer's instructions (Thermo Fisher Scientific). Primer sequences for *marA* and *recA* were as previously described²⁹. All experiments were conducted in at least three biological replicates, and relative expression levels were estimated as previously described⁴³, using *recA* expression as reference.

Immunoblot analysis for protein quantification. Overnight cultures of *E. coli* BW25113 and the *marA* deletion mutant ($\Delta marA$) were diluted 1:1,000 into 50 ml LB containing 500 µg/ml aspirin, 150 µg/ml vanillin or DMSO (solvent control), followed by growth with agitation at 37 °C to OD_{595 nm} ~ 0.5. Cells were washed

in PBS containing corresponding drugs or DMSO, then resuspended to match OD_{595 nm} = 1. Cell pellets were resuspended in Laemmli buffer and heated to 95 °C for 3 min followed by immunoblot analysis with anti-AcrA polyclonal antiserum (gift from K. M. Pos) at 1:200,000 dilution. Primary antiserum was detected using anti-rabbit HRP (A0545 Sigma) at 1:5,000 dilution. Cell loading was controlled with the anti-RecA antibody (rabbit, ab63797 Abcam). For densitometry analysis, the pixel intensity of AcrA bands from cell-density-normalized samples was quantified using ImageJ. At least four different biological replicates were blotted. Each biological replicate was run and blotted twice (technical replicates). Relative AcrA levels per biological replicate correspond to the average intensities of the technical replicates. All blots can be found in Supplementary Fig. 1.

Screening the *E. coli* Keio knockout collection for identifying the mode of action of drug interactions.

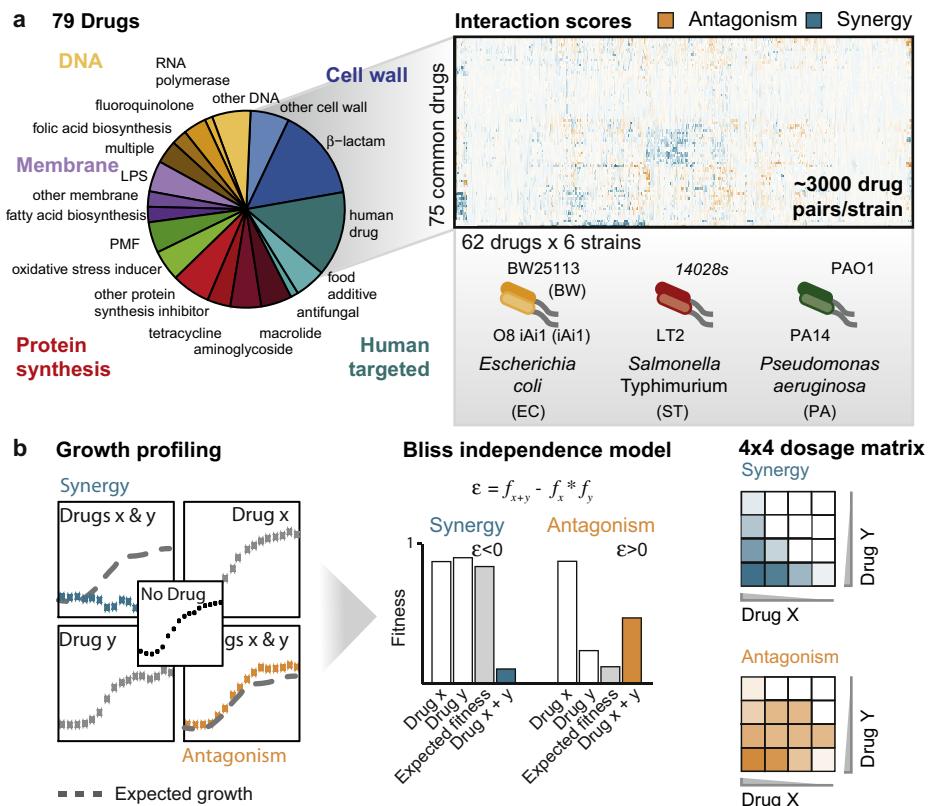
The *E. coli* Keio knockout collection²⁶ (two independent clones per mutant) was arrayed in 1,536-format on LB agar plates using a Rotor HDA (Singer Instruments) as previously described²⁹. The growth of each mutant was estimated by colony opacity⁴⁴ after a 13-h incubation at 37 °C in the absence and presence of vanillin (200 µg/ml), spectinomycin (4 µg/ml) and their combination. All plates were imaged under controlled lighting conditions (spImager, S&P Robotics) using an 18-megapixel Canon Rebel T3i (Canon). Experiments were done in biological triplicates. The fitness of each mutant was calculated by dividing the growth in condition (vanillin, spectinomycin or both) by the growth in LB, after correcting for outer-frame plate effects⁴⁴. Bliss scores were calculated according to equation (1) per replicate and then averaged (Supplementary Table 7).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The code used for data analysis is available from <https://github.com/AnaRitaBrochado/DrugInteractionsPipeline>.

Data availability. All data supporting the findings of this study are included in this paper as Supplementary Information.

33. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
34. Saka, K. et al. A complete set of *Escherichia coli* open reading frames in mobile plasmids facilitating genetic studies. *DNA Res.* **12**, 63–68 (2005).
35. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
36. Yeh, P. J., Hegreness, M. J., Aiden, A. P. & Kishony, R. Drug interactions and the evolution of antibiotic resistance. *Nat. Rev. Microbiol.* **7**, 460–466 (2009).
37. Bliss, C. I. The toxicity of poisons applied jointly. *Ann. Appl. Biol.* **26**, 585–615 (1939).
38. Loewe, S. Die quantitativen Probleme der Pharmakologie. *Ergeb. Physiol.* **27**, 47–187 (1928).
39. Szappanos, B. et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* **43**, 656–662 (2011).
40. Mateus, A. et al. Prediction of intracellular exposure bridges the gap between target- and cell-based drug discovery. *Proc. Natl Acad. Sci. USA* **114**, E6231–E6239 (2017).
41. Richter, M. F. et al. Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature* **545**, 299–304 (2017).
42. Piddock, L. J., Jin, Y. F., Ricci, V. & Asuquo, A. E. Quinolone accumulation by *Pseudomonas aeruginosa*, *Staphylococcus aureus* and *Escherichia coli*. *J. Antimicrob. Chemother.* **43**, 61–70 (1999).
43. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_t}$ method. *Methods* **25**, 402–408 (2001).
44. Kritikos, G. et al. A tool named Iris for versatile high-throughput phenotyping in microorganisms. *Nat. Microbiol.* **2**, 17014 (2017).
45. Safdar, N., Handelsman, J. & Maki, D. G. Does combination antimicrobial therapy reduce mortality in Gram-negative bacteraemia? A meta-analysis. *Lancet Infect. Dis.* **4**, 519–527 (2004).
46. Taber, H. W., Mueller, J. P., Miller, P. F. & Arrow, A. S. Bacterial uptake of aminoglycoside antibiotics. *Microbiol. Rev.* **51**, 439–457 (1987).
47. Mazzariol, A., Tokue, Y., Kanegawa, T. M., Cornaglia, G. & Nikaido, H. High-level fluoroquinolone-resistant clinical isolates of *Escherichia coli* overproduce multidrug efflux protein AcrA. *Antimicrob. Agents Chemother.* **44**, 3441–3443 (2000).
48. Davis, B. D., Chen, L. L. & Tai, P. C. Misread protein creates membrane channels: an essential step in the bactericidal action of aminoglycosides. *Proc. Natl Acad. Sci. USA* **83**, 6164–6168 (1986).
49. Miller, P. F., Gambino, L. F., Sulavik, M. C. & Gracheck, S. J. Genetic relationship between *soxRS* and *mar* loci in promoting multiple antibiotic resistance in *Escherichia coli*. *Antimicrob. Agents Chemother.* **38**, 1773–1779 (1994).
50. Fernandes, F., Neves, P., Gameiro, P., Loura, L. M. & Prieto, M. Ciprofloxacin interactions with bacterial protein OmpF: modelling of FRET from a multi-tryptophan protein trimer. *Biochim. Biophys. Acta* **1768**, 2822–2830 (2007).
51. Machado, D. et al. Ion channel blockers as antimicrobial agents, efflux inhibitors, and enhancers of macrophage killing activity against drug resistant *Mycobacterium tuberculosis*. *PLoS ONE* **11**, e0149326 (2016).
52. Maier, L. et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).

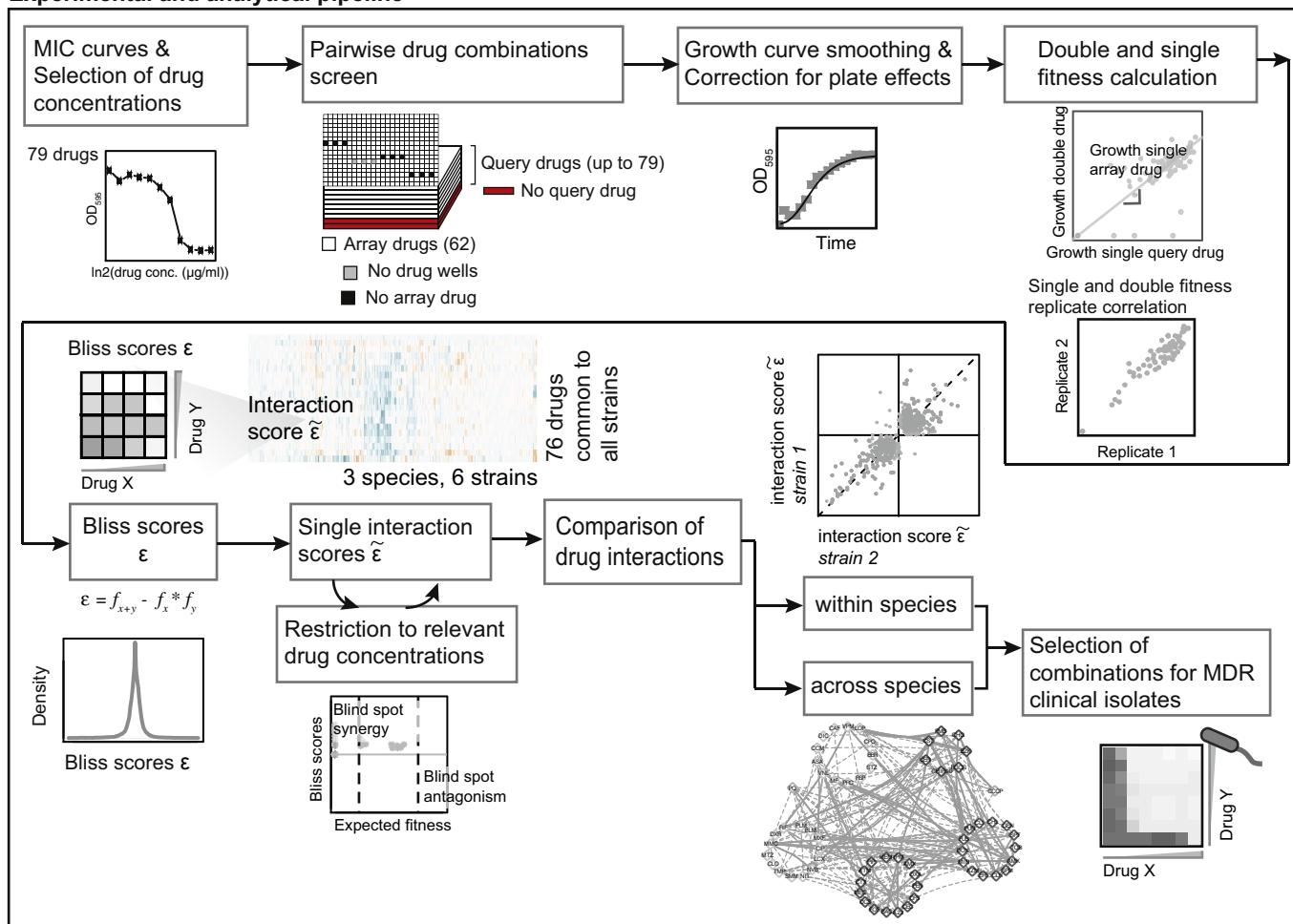


Extended Data Fig. 1 | High-throughput profiling of pairwise drug combinations in Gram-negative bacteria. **a**, Drug and species selection for screen. The 79 drugs used in the combinatorial screen are grouped according to categories (Supplementary Table 1). Antibacterial agents are grouped by target, with the exception of antibiotic classes for which enough representatives were screened (>2) to form a separate category (β -lactams, macrolides, tetracyclines, fluoroquinolones and aminoglycosides). Classification of human-targeted drugs and food additives is not further refined, because for most of these the mode of action is unclear. A subset of 62 arrayed drugs was profiled against

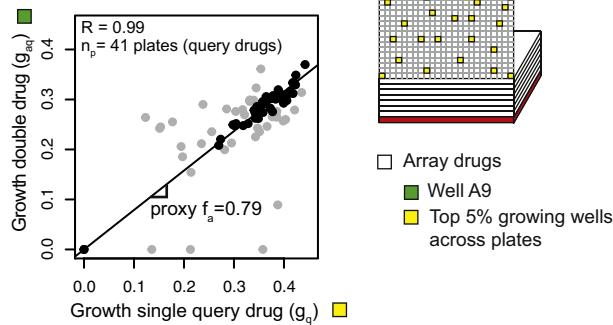
79 drugs in all 6 strains (75 drugs were common to all strains and are depicted in the heat map). Strains are colour-coded according to species: yellow, *E. coli*; red, *S. Typhimurium*; green, *P. aeruginosa*. **b**, Quantification of drug–drug interactions. Growth was profiled by measuring optical density ($OD_{595\text{ nm}}$) over time in the presence of no, one and both drugs. x and y correspond to particular concentrations of drugs X and Y . Interactions were defined according to Bliss independence. Significantly lower or higher fitness than the expectation ($f_x \times f_y$) indicates synergy or antagonism, respectively. Synergy and antagonism were assessed by growth in 4×4 checkerboards (Methods).

a

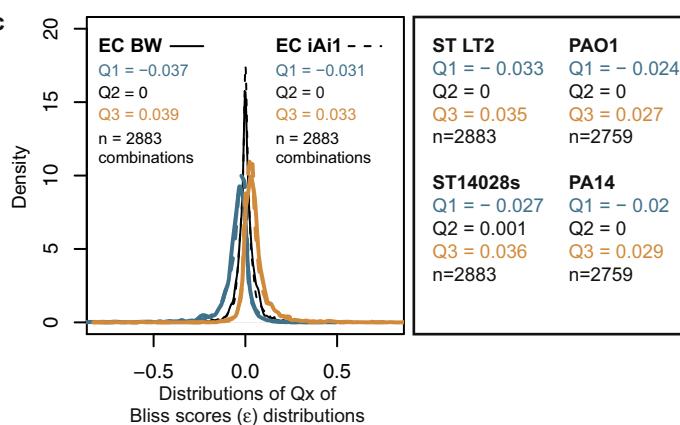
Experimental and analytical pipeline



b Well A9: EC BW

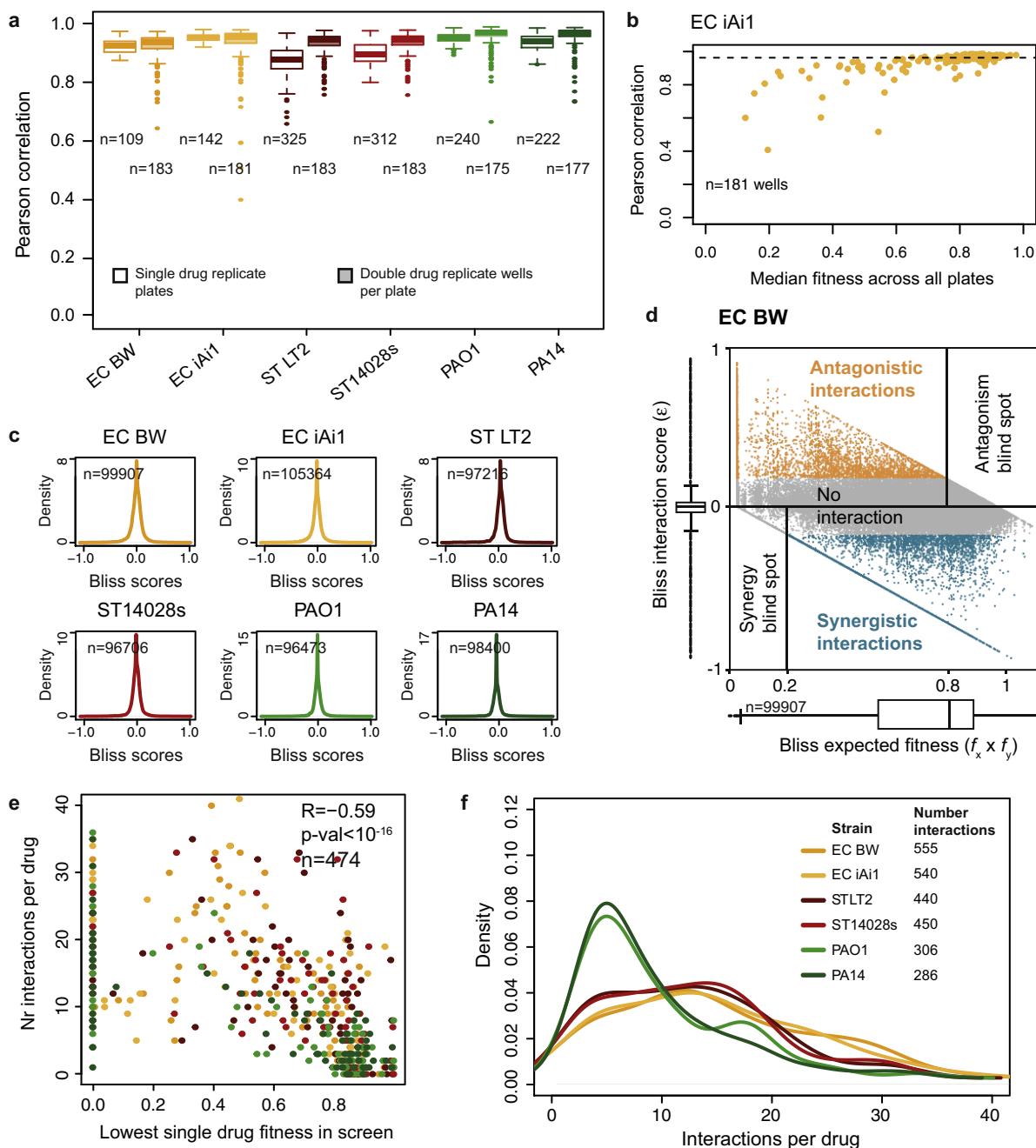
Array drug: spectinomycin 3 $\mu\text{g/ml}$ 

c



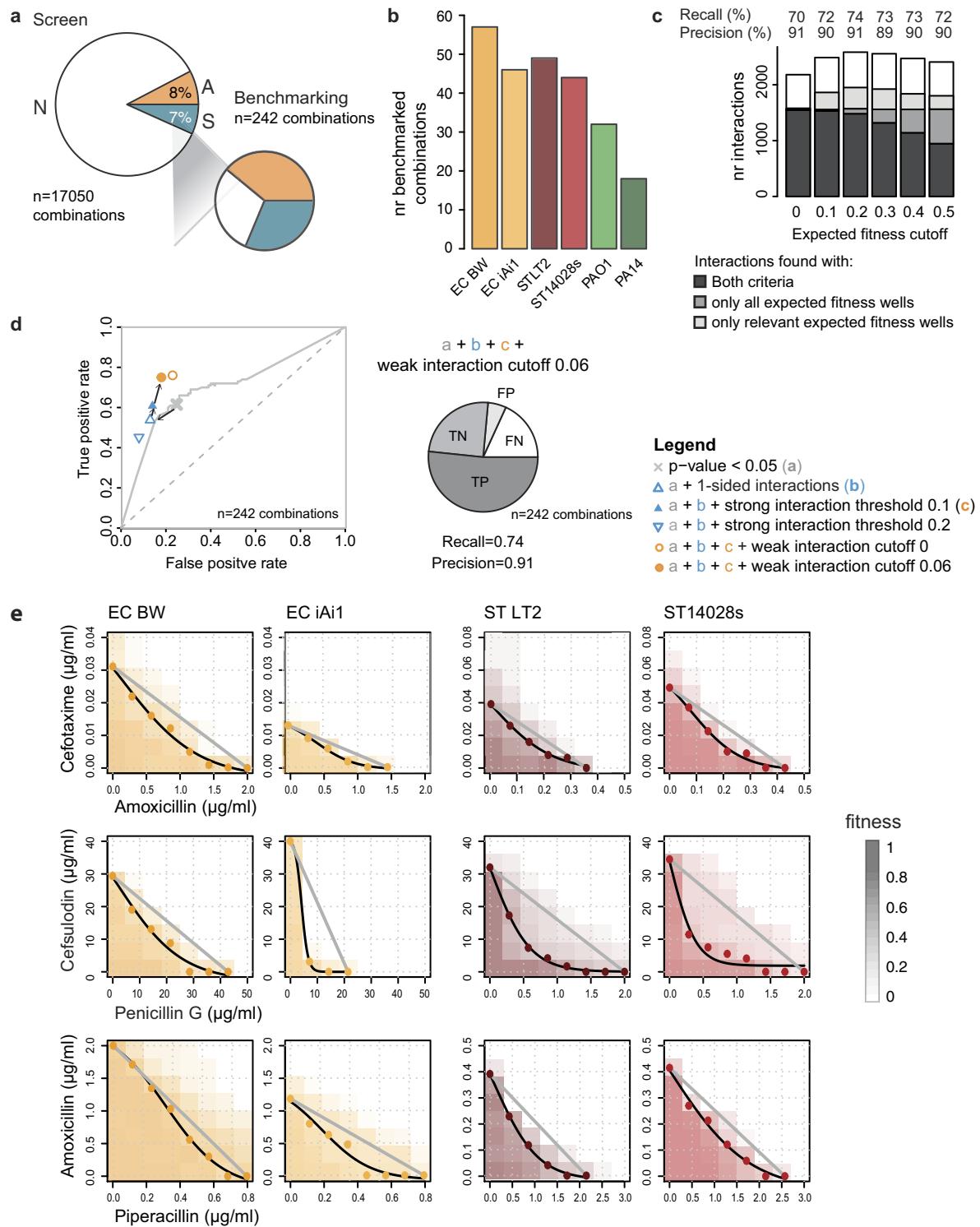
Extended Data Fig. 2 | Data analysis pipeline. a, Flowchart of the data analysis pipeline. b, Estimating single-drug fitness of arrayed drugs. As drug-drug interactions are rare, the slope of the line of best fit between g_{aq} (growth with double drug) and g_q (growth with query drug alone, which was deduced from the average of the top 5% growing wells across plates within a batch), across plates (n_p) of query drugs within a batch, corresponds to a proxy of the fitness of the arrayed drug alone, f_a (see Methods). r denotes the Pearson correlation coefficient between g_{aq} and g_q across n_p . Well A9 from *E. coli* BW25113, containing 3 $\mu\text{g ml}^{-1}$

spectinomycin, is shown as an example of arrayed drugs with several interactions. Several query drugs deviated from the expected fitness (light grey points), and therefore only half of the plates corresponding to the interquartile range of g_{aq}/g_q were used to estimate f_a . c, Density distributions of the first, second and third quartiles of Bliss-score (ε) distributions for *E. coli*. Q_1 , Q_2 and Q_3 denote the median of the first, second and third quartiles of ε distributions, respectively. n denotes the number of drug combinations used.



Extended Data Fig. 3 | Data quality control. **a**, High replicate correlation for single- and double-drug treatments. Transparent box plots contain Pearson correlation coefficients between plates of the same batch that contain only arrayed drugs (for which LB was used instead of the second drug). n represents the total number of correlations. Full box plots contain Pearson correlation coefficients between double-drug replicate wells within the same plate, across all plates. n represents the number of wells used for correlation, $n_{\max} = (62 \text{ drugs} + 1 \text{ LB}) \times 3 \text{ concentrations} = 189$. Only wells with median growth above 0.1 were taken into account for this correlation analysis (see **b**). For all box plots the centre line, limits, whiskers and points correspond to the median, upper and lower quartiles, $1.5 \times$ interquartile range and outliers, respectively. **b**, Wells with lower median growth have lower replicate correlation. The double-drug correlation coefficients used to generate the box plot from **a** are plotted as a function of the median growth of all wells across all plates for *E. coli* iAi1. Wells with overall lower growth (due to the strong inhibition of the arrayed drug) are less reproducible owing to a combination of the lower spread of growth values and the sigmoidal nature of the drug–dose response curves. **c**, Drug–drug interactions are rare. Density distributions

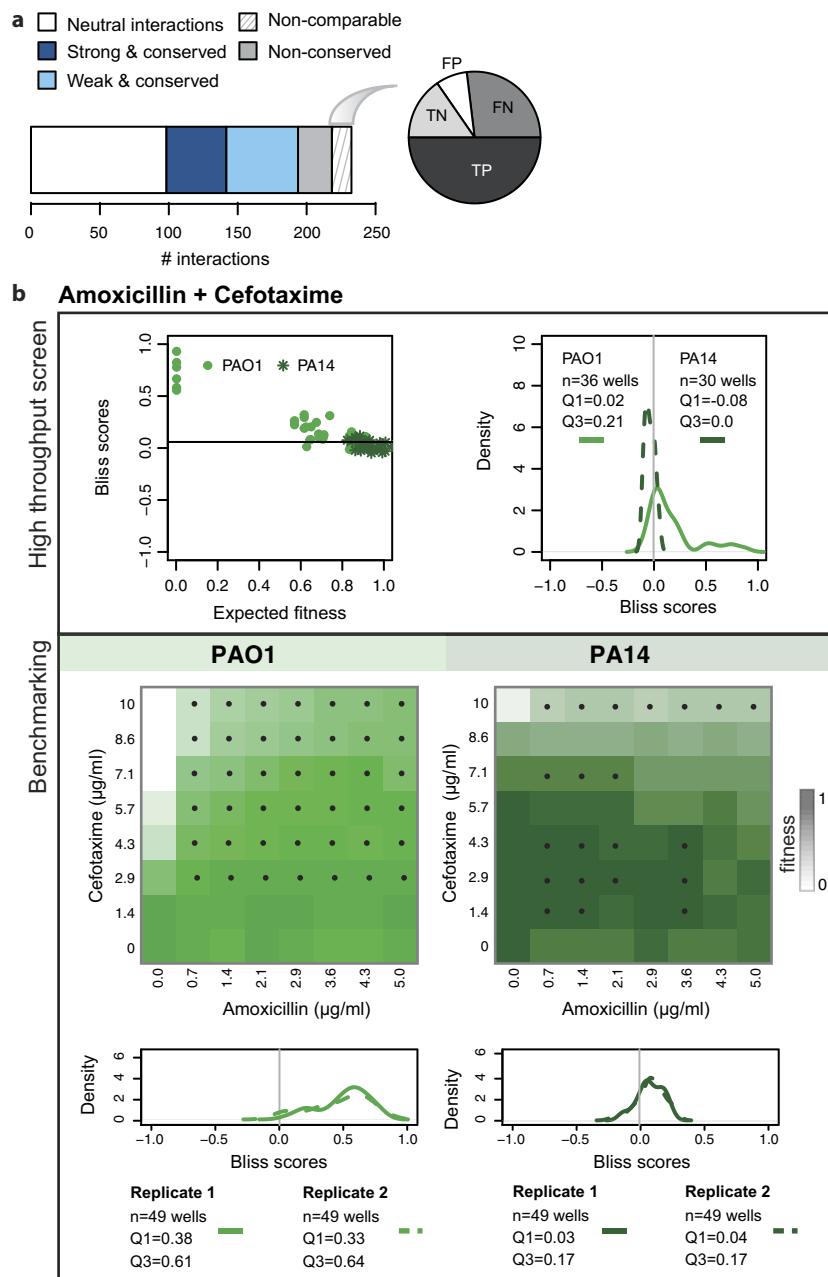
of all Bliss scores (ε) obtained per strain. **d**, The ability to detect synergies and antagonisms depends on the effects of single-drug treatments. Bliss scores (ε) are plotted as function of expected fitness ($f_x \times f_y$) for all drug concentration ratios for all combinations in *E. coli* BW25113 (as an example). Box plots summarizing both variables are shown besides the axes ($n = 99,907$ Bliss scores; centre line, limits, whiskers and points correspond to the median, upper and lower quartiles, $1.5 \times$ interquartile range and outliers, respectively). Blind spots for detecting antagonism and synergy are indicated; both of these are based on the expected fitness (see also Extended Data Fig. 4c, d), and are therefore dependent on the growth of the strain with the single drugs. The number of drug combinations falling in the blind spot for antagonism is larger, owing to the number of drugs used in the screen that do not inhibit *E. coli* on their own. **e**, Scatter plot of the number of interactions per drug versus the minimum fitness of the drug alone (as obtained in the screen, Supplementary Table 1). Strong and weak interactions are represented. n denotes the total number of interactions and r is the Pearson correlation coefficient. Strains are colour-coded as above. **f**, Density distributions of the number of interactions per drug for all strains.



Extended Data Fig. 4 | See next page for caption.

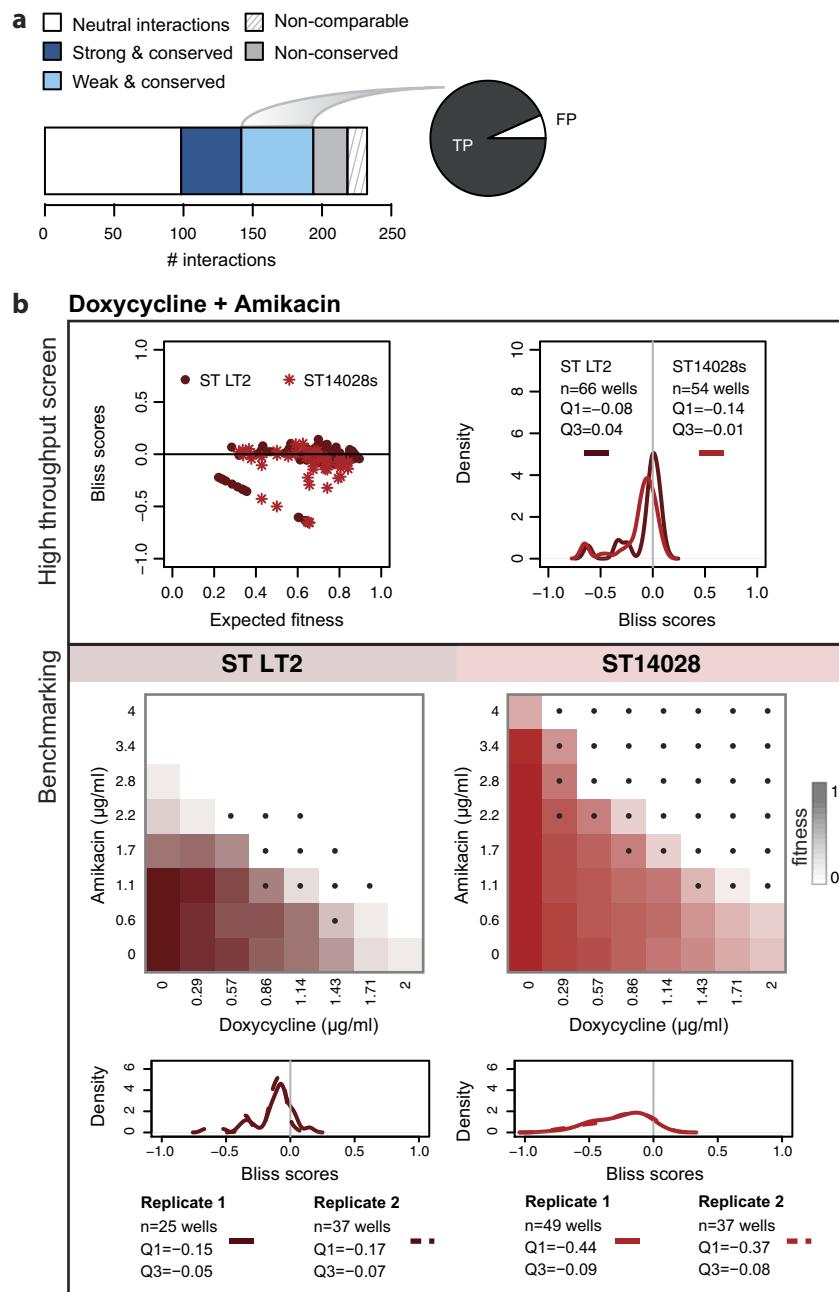
Extended Data Fig. 4 | Benchmarking and sensitivity analysis. **a**, The validation set is enriched in synergies and antagonisms to better assess true and false positives. Comparison of percentages of synergy and antagonism between the screen and validation set. Both strong and weak interactions (Fig. 2b) are accounted for in the screen tally. **b**, Number of benchmarked interactions per strain. **c, d**, Sensitivity analysis of the statistical thresholds for calling interactions. **c**, The total number of interactions as a function of the expected fitness ($f_x \times f_y$) cutoff was used to restrict the ε distributions to relevant drug concentrations. Strong drug–drug interactions are classified according to the ε distribution in which they were significant: complete distribution only (that is, all expected fitness wells), relevant wells only (that is, all wells with $f_x \times f_y >$ cutoff for synergies and all wells with $f_x \times f_y < (1 - \text{cutoff})$ for antagonisms), or in both. Weak drug–drug interactions are independently assigned and represented in white. We selected an expected fitness cutoff of 0.2, as this cutoff resulted in the largest number of total interactions detected, with the highest precision and recall (91 and 74%, respectively) after benchmarking against the validation dataset. **d**, Receiver operating characteristic curve for the screen across different P value thresholds (10,000 repetitions of a two-sided permutation test of Wilcoxon rank-sum test after correction for multiple

testing, see Methods) as a unique criterion for assigning interactions. The selected P value (0.05) for the screen threshold is indicated by a grey cross. Sensitivity to additional parameters for calling hits is shown: allowing interactions to be either antagonisms or synergies but not both (one-sided); as well as strong and weak interaction thresholds. True- and false-positive rates were estimated based on the validation dataset. Precision and recall for the final and best-performing set of parameters are shown: one-sided interactions, $P < 0.05$, $f_x \times f_y$ cutoff = 0.2 and $|\varepsilon| > 0.1$ for strong interactions, $|\varepsilon| > 0.06$ for weak interactions. TP, true positive; TN, true negative; FP, false positive; FN, false negative. n indicates the total number of benchmarked drug combinations (Supplementary Table 3). **e**, Synergies between β -lactams according to the Loewe additivity interaction model. The results of 8×8 checkerboards for 3 combinations between β -lactams in 4 strains are shown. The grey line in each plot represents the null hypothesis in the Loewe additivity model and the black line corresponds to the IC_{50} isobole, which was estimated by fitting a logistic curve to the interpolated drug concentrations (coloured dots, Methods). Piperacillin did not reach 50% growth inhibition in *E. coli*, thus IC_{20} and IC_{40} isoboles were used for the amoxicillin + piperacillin combination in *E. coli* BW25113 and *E. coli* IAI1, respectively.



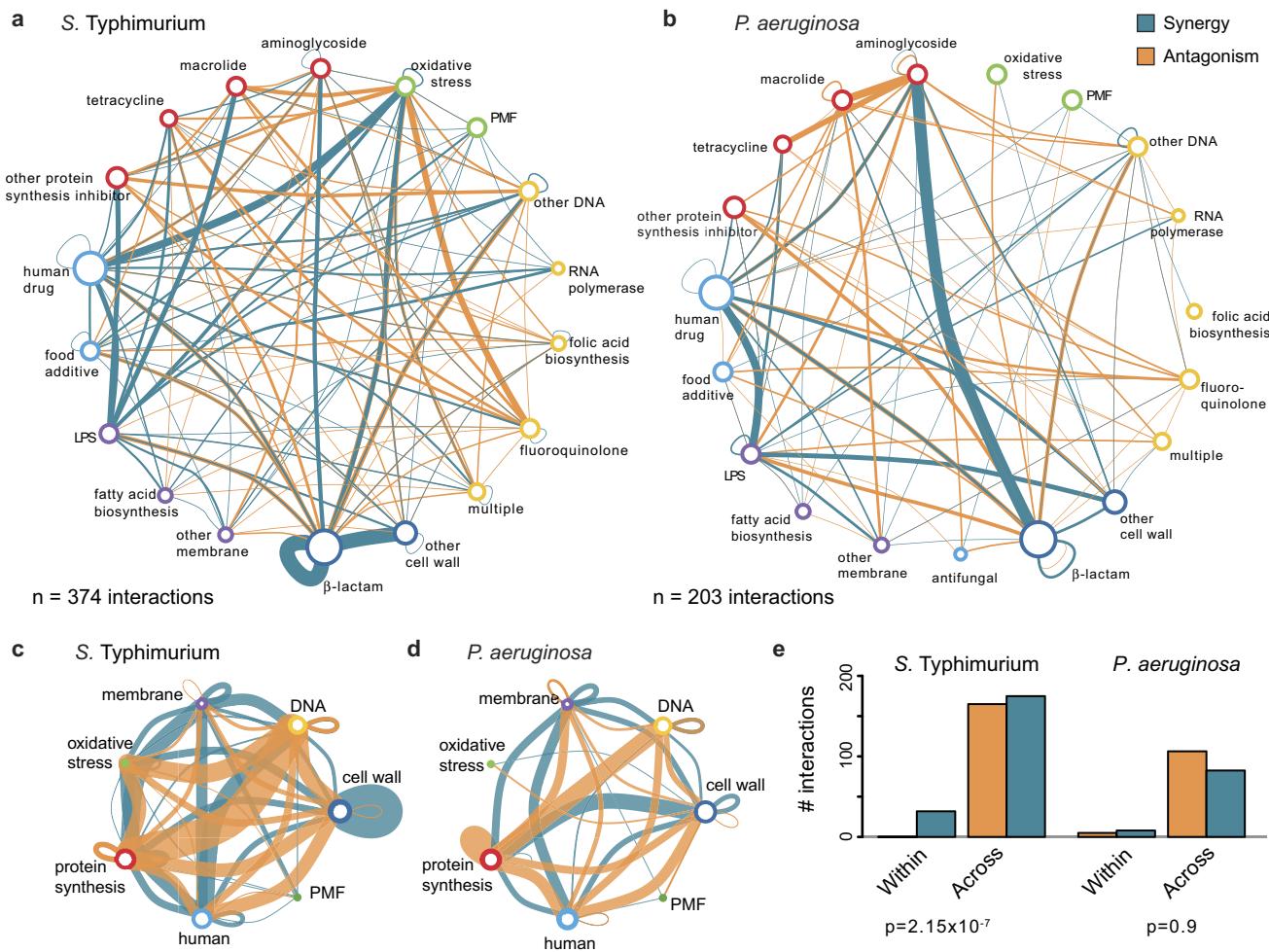
Extended Data Fig. 5 | Benchmarking of non-comparable drug–drug interactions. **a**, The bar plot illustrates the division of benchmarked drug combinations according to their degree of conservation within species. The pie chart shows the proportion of false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN) within non-comparable drug–drug interactions. **b**, Combination of amoxicillin with cefotaxime in *P. aeruginosa* as an example of a non-comparable drug–drug interaction. Top box, the results of the screen. Left, Bliss scores as function of expected fitness for both strains. Right, a density distribution

of the Bliss scores. n denotes the total number of Bliss scores, Q1 and Q3 indicate the Bliss score for the first and third quartiles, respectively. Antagonism was detected only for PAO1 ($Q3 > 0.1$). PA14 was resistant to both drugs at concentrations screened (top left panel), rendering the detection of antagonism impossible. Bottom box, benchmarking results indicate that the interaction is antagonistic in both strains, albeit weaker in PA14 and visible mostly at higher concentrations. The colour intensity on checkerboard reflects fitness and black dots correspond to drug ratios in which the Bliss score is above 0.1.



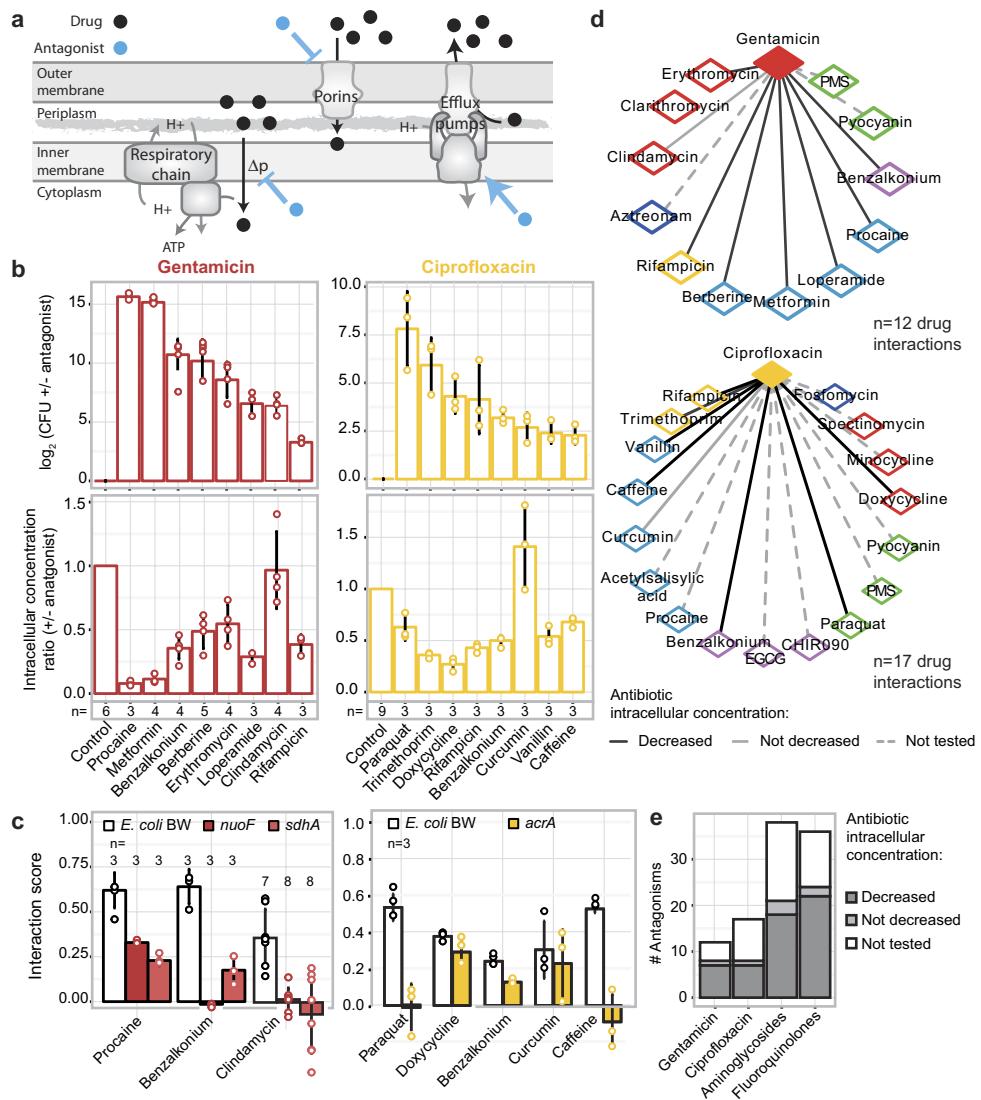
Extended Data Fig. 6 | Benchmarking of weak conserved drug–drug interactions. **a**, The bar plot illustrates the division of benchmarked drug combinations as in Extended Data Fig. 5a. The pie chart shows the proportion false positives and true positives within weak conserved interactions. **b**, Combination of doxycycline with amikacin in *S. Typhimurium* as an example of a weak conserved drug–drug interaction. Top box, the results of the screen. Left, Bliss scores as a function of expected fitness for both strains. Right, a density distribution of the Bliss

scores. n denotes the total number of Bliss scores, Q1 and Q3 indicate the Bliss score for quartiles 1 and 3, respectively. A strong synergy was detected only for ST14028 ($Q1 < -0.1$), and a weak conserved synergy was assigned afterwards to ST LT2 ($Q1 < -0.06$). Bottom box, the benchmarking results confirm that the interaction is synergistic in both strains. The colour intensity on checkerboard reflects fitness and black dots correspond to drug ratios in which the Bliss score is below -0.1 .



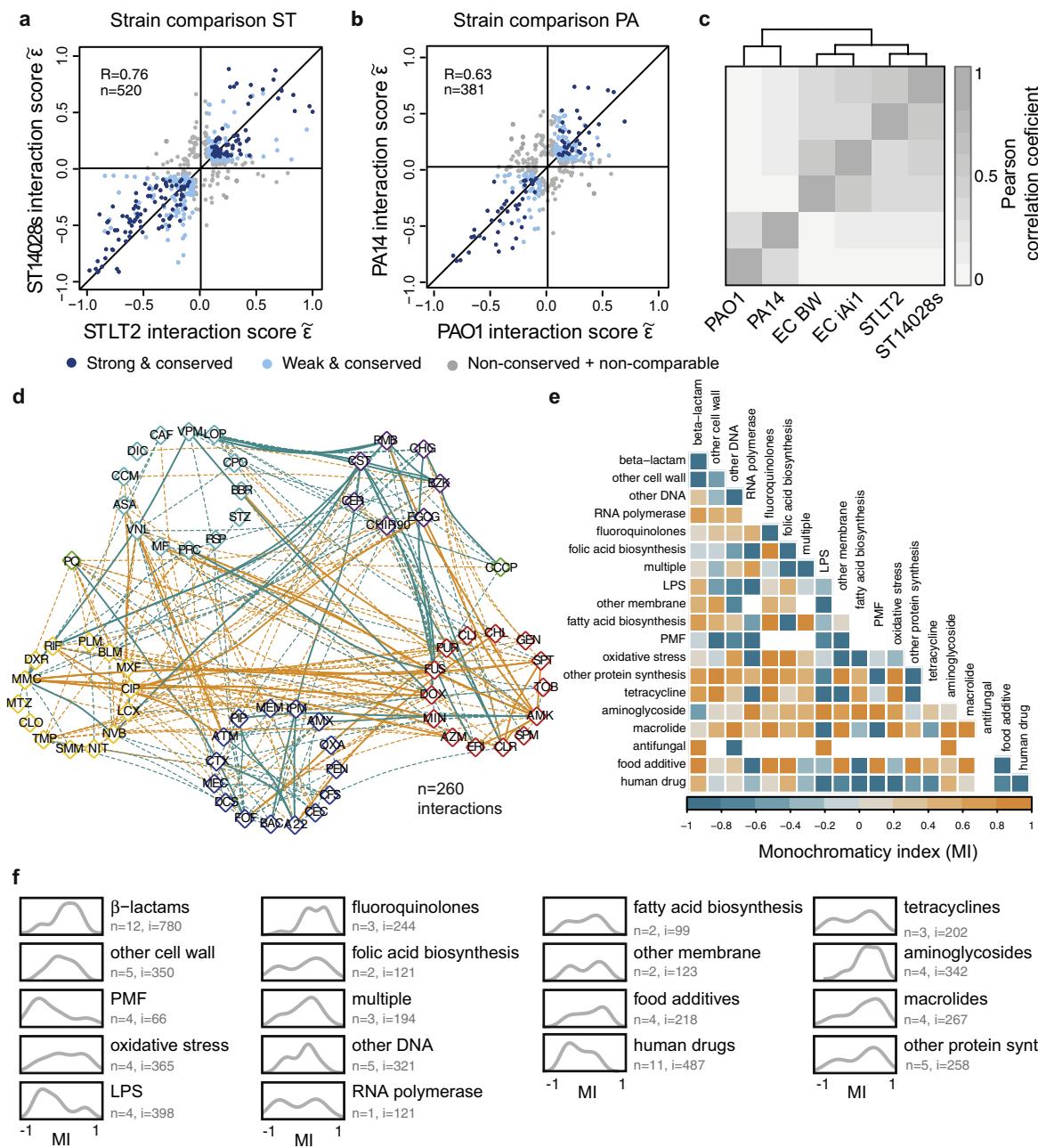
Extended Data Fig. 7 | *Salmonella* and *Pseudomonas* drug–drug interaction networks. **a, b**, Drug category interaction networks. Nodes represent drug categories according to Extended Data Fig. 1a, and plotted as in Fig. 1b. Conserved interactions, including weak conserved interactions, are shown here. One of the most well-known and broadly used synergies is that of aminoglycosides and β -lactams⁴⁵. Consistent with its use against *P. aeruginosa* in clinics, we detected multiple strong synergies between specific members of the two antibiotic classes in *P. aeruginosa* but fewer interactions in the other two species.

c, d, Drug–drug interactions across cellular processes. Representation as in **a, b** but grouping drug categories targeting the same general cellular process. **e**, Quantification of synergy and antagonism in the networks from **a, b** and the corresponding χ^2 -test *P* value. As in *E. coli* (Fig. 1), antagonism occurs more frequently than synergy and almost exclusively between drugs belonging to different categories in *S. Typhimurium* and *P. aeruginosa*. In *P. aeruginosa*, there are very few interactions occurring between drugs of the same category (within the group).



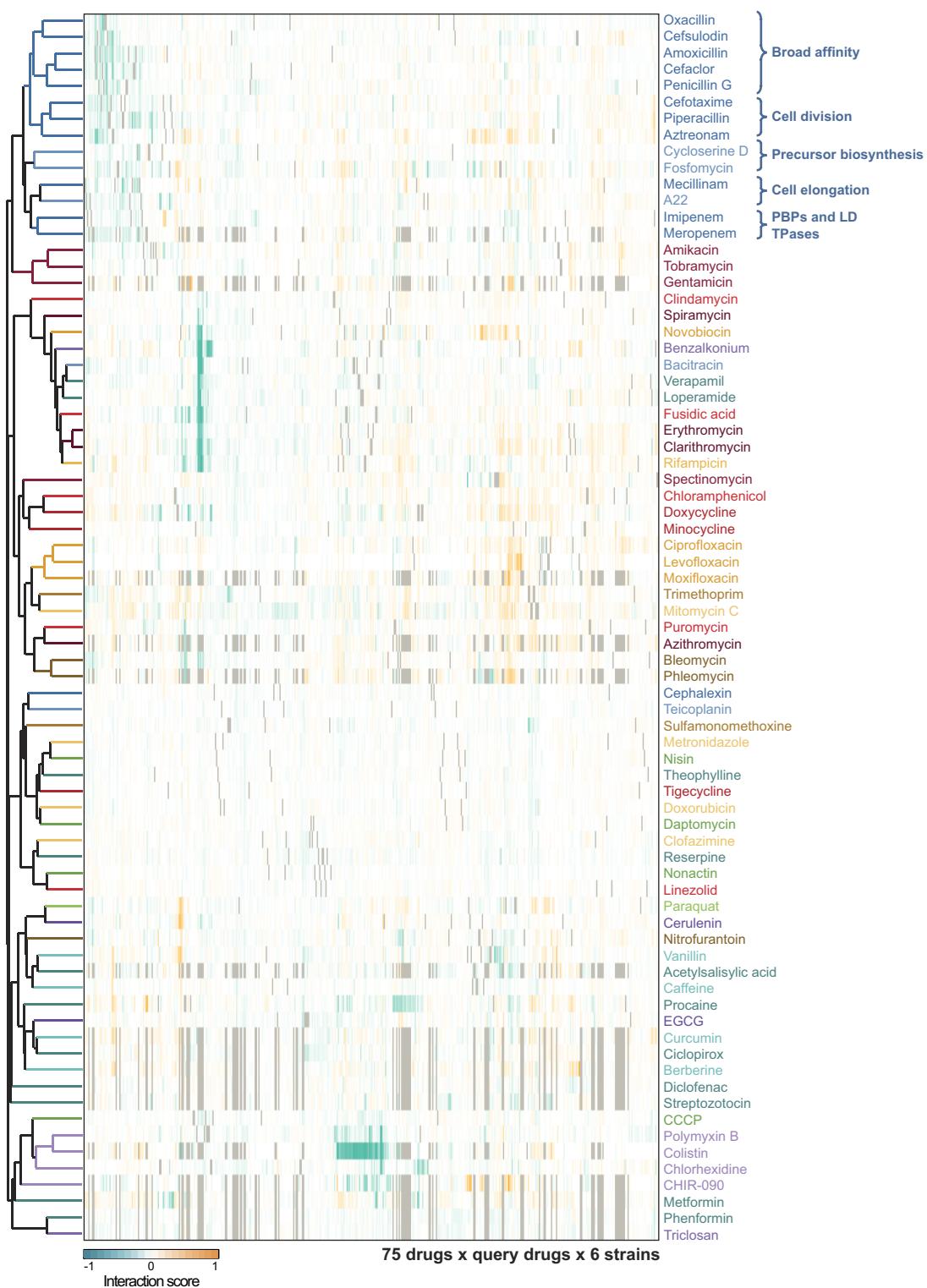
Extended Data Fig. 8 | Drug antagonisms are often due to a decrease in intracellular drug concentrations. **a**, Cartoon of possible modes of action for drug–drug interactions that function via modulation of the intracellular drug concentration. A given drug (antagonist, blue) inhibits the uptake or promotes the efflux of another drug (black), and thus decreases its intracellular concentration. **b**, Different antagonists (see Methods for concentrations) of gentamicin (red, 5 $\mu\text{g ml}^{-1}$) and ciprofloxacin (yellow, 2.5 $\mu\text{g ml}^{-1}$) identified in our screen for *E. coli* BW25113 also rescue the killing effect of the two bactericidal drugs in the same strain, or its parental MG1655 (top right and top left panels, respectively). With the exception of clindamycin (for gentamicin) and curcumin (for ciprofloxacin), all other antagonists decrease the intracellular concentration of their interacting drug (bottom panels). Gentamicin was detected by using radiolabelled compound, and ciprofloxacin with LC–MS/MS (Methods). The degree of rescue (top panels) in many cases follows the decrease in intracellular concentration (bottom panels), which implies that most of these interactions depend at least partially on modulating the intracellular concentration of the antagonized drug. **c**, Antagonisms are resolved in *E. coli* BW25113 mutants that lack key components that control the intracellular concentration of the antagonized drug. Aminoglycosides depend on proton motive force-energized uptake, and thus on respiratory complexes^{7,46}; ciprofloxacin is effluxed by AcrAB–TolC^{29,47}. For gentamicin, most interactions are resolved when respiration is defected, even the interaction with clindamycin (which does not modulate intracellular gentamicin concentration, see **b**); this presumably occurs because the mode of action and import of aminoglycosides are linked by a positive feedback

loop^{7,48}. For ciprofloxacin, antagonisms with paraquat and caffeine are resolved in the $\Delta acrA$ mutant, which implies that both compounds induce the AcrAB–TolC pump (well-established for paraquat⁴⁹). By contrast, interactions with curcumin, benzalkonium and doxycycline remain largely intact in the $\Delta acrA$ mutant. The first interaction is expected, as curcumin does not modulate intracellular ciprofloxacin concentration (see **b**). In the other two cases, other component(s) besides AcrAB–TolC may be responsible for the altered ciprofloxacin import and/or export; for example, ciprofloxacin uses OmpF to enter the cell⁵⁰. Ciprofloxacin and gentamicin concentrations were adjusted in all strains according to MIC (70% and 100% MIC for ciprofloxacin and gentamicin, respectively; all drug concentrations are listed in Supplementary Table 6). Bliss interaction scores (ε) were calculated as in the screen. Bar plots and error bars in **b**, **c** represent the average and s.d., respectively, across n independent biological replicates. **d**, Gentamicin and ciprofloxacin antagonism networks for *E. coli* BW. Nodes represent drugs coloured according to targeted cellular process (as in Extended Data Fig. 1a). Full and dashed edges represent antagonistic drug–drug interactions for which intracellular antibiotic concentration was and was not measured, respectively. Drug interactions that result in decreased intracellular concentration of the antagonized drug are represented by black edges. **e**, Quantification of antagonistic drug–drug interactions from the networks in **d**. The bars for fluoroquinolones and aminoglycosides account for an extrapolation of antagonistic interactions to all other members of the two classes, assuming that they behave in the same way as ciprofloxacin and gentamicin, respectively.



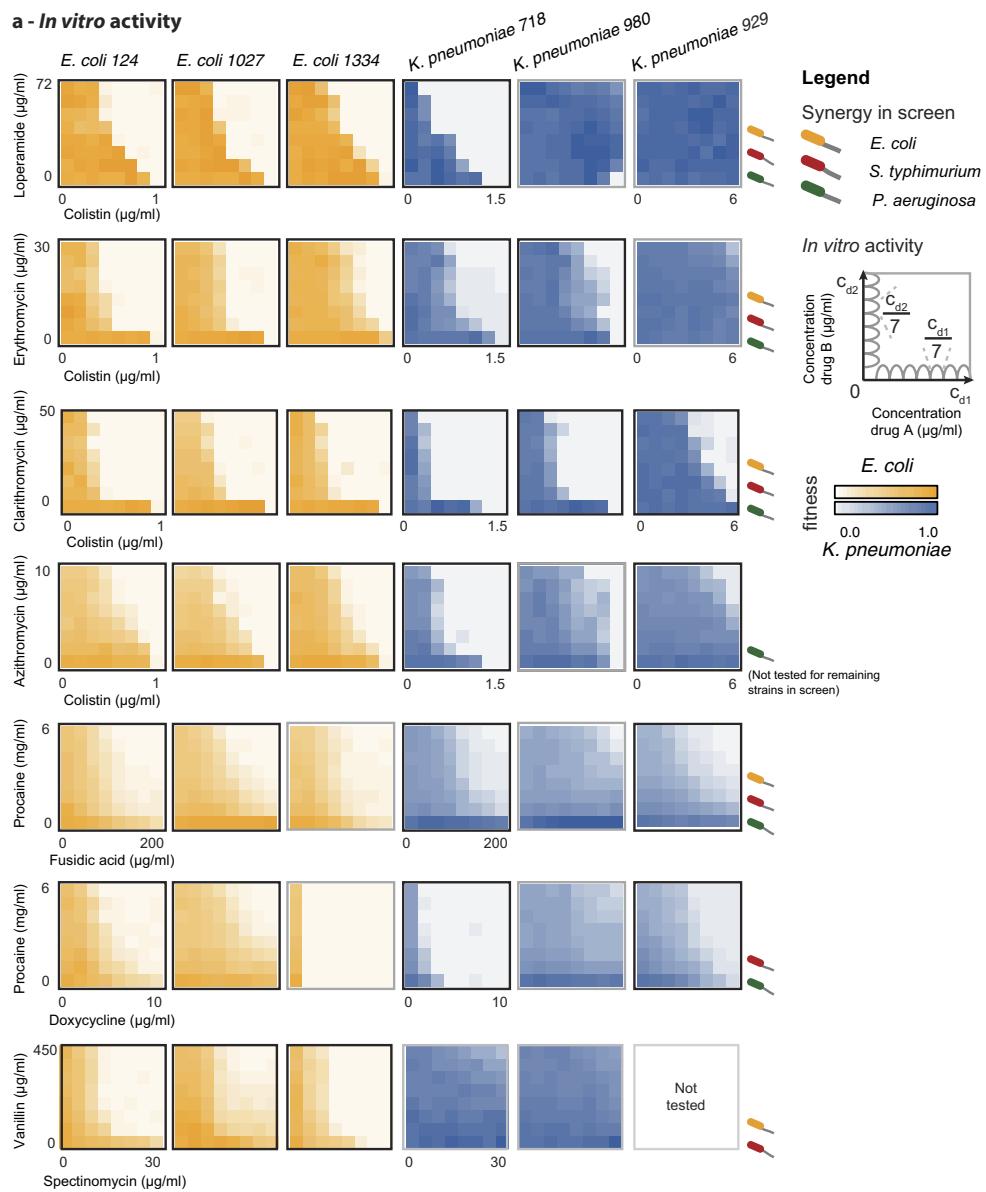
Extended Data Fig. 9 | Drug–drug interactions are largely conserved within species and only partially driven by mode of action. **a, b**, Drug–drug interactions are conserved in *S. Typhimurium* (a) and *P. aeruginosa* (b). Scatter plot of interaction scores in the two strains of each species; only strong interactions for at least one strain are shown. Colours and grouping as in Fig. 2a. r denotes the Pearson correlation and n denotes the total number interactions plotted. The lower correlation in *P. aeruginosa* is presumably due to fewer and weaker interactions. **c**, Drug interaction profiles are driven by phylogeny. Clustering of strains based on the Pearson correlation of their drug interaction profiles (taking into account all pairwise drug combinations; $n = 2,759$ –2,883 depending on the species). Strains of the same species cluster together; the two enterobacterial species—*E. coli* and *S. Typhimurium*—behave more similarly to one another than either does to the phylogenetically more-distant *P. aeruginosa*. **d**, Conserved drug–drug interaction network. Nodes represent individual drugs grouped and coloured by targeted cellular process (as in Extended Data Fig. 1a). Drug names are represented by three-letter codes (given in Supplementary Table 1). Dashed and full edges correspond to conserved interactions between two or three species, respectively. Many of the human-targeted drugs, such as loperamide,

verapamil and procaine, exhibit a general potentiating effect that is similar to that of membrane-targeting drugs. This suggests that these drugs may also facilitate drug uptake or impair efflux, consistent with previous reports on the role of loperamide in *E. coli* and verapamil in *Mycobacterium tuberculosis*^{4,51}. **e**, Monochromativity between all drug categories. The monochromativity index (MI) reflects whether interactions between drugs of two categories are more synergistic (MI = −1) or antagonistic (MI = 1) than the background proportion of synergy and antagonism. The MI equals zero when interactions between two drug categories have the same proportion of synergy and antagonism as all interactions together (Methods). The MI was calculated using all interactions from the six strains for all category pairs that had at least two interactions. White cells in the heat map correspond to category pairs for which no (or an insufficient number of) interactions were observed. **f**, Human-targeted drugs, and LPS or proton motive force inhibitors, are strong and promiscuous adjuvants. Density distributions of the monochromativity indices per drug category from e are shown. n denotes the number of drugs in each category and i the number of interactions in which they are involved.



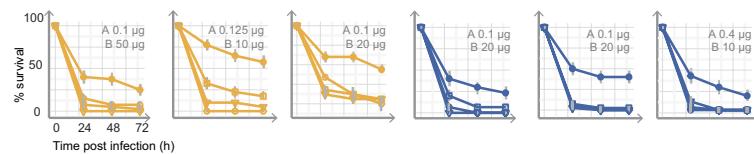
Extended Data Fig. 10 | Hierarchical clustering of drugs according to their interaction profiles. Rows depict the 75 drugs common to all strains (coloured according to drug category, see Extended Data Fig. 1a), and

columns depict their interactions with other drugs in all six strains tested. Clustering was done using the median of the ε distributions, uncentred correlation and average linkage.

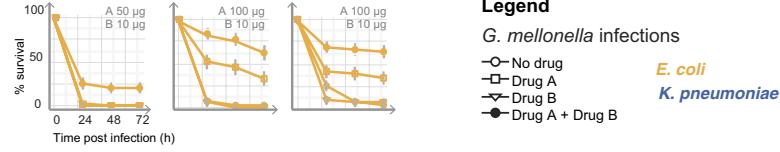


b - *G. mellonella* infections

Drug A: Colistin Drug B: Clarithromycin



Drug A: Spectinomycin Drug B: Vanillin



Legend

G. mellonella infections

- No drug
- Drug A
- △ Drug B
- Drug A + Drug B

E. coli

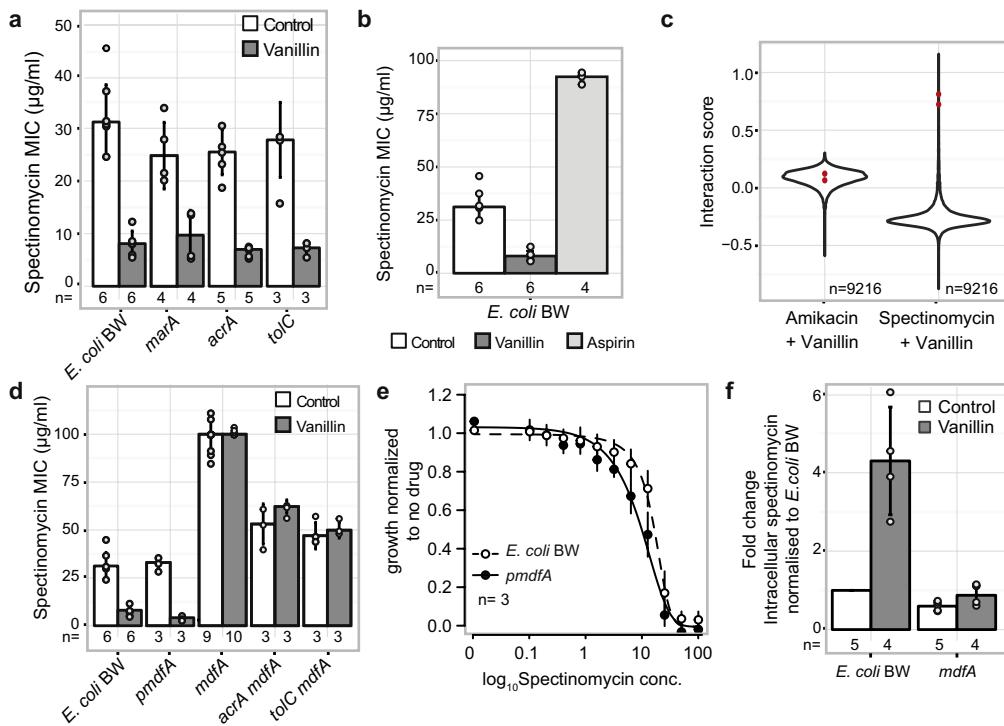
K. pneumoniae

Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Active synergies against Gram-negative MDR clinical isolates in vitro and in the *G. mellonella* infection model.

Both human-targeted drugs (which have recently been found to have an extended effect on bacteria⁵²) and food additives can promote the action of antibiotics in MDR strains, indicating that their use as antibacterial adjuvants should be explored further. **a**, Drug combinations active against MDR *E. coli* and *K. pneumoniae* clinical isolates (see also Fig. 4). Interactions are shown as 8×8 checkerboards and synergies have a black bold border. Drug pairs are the same for each row of panel **a**, and are indicated at the first checkerboard in each row. The species in which the interaction was detected in the screen are indicated after the last

checkerboard in each row. Concentrations increase in equal steps per drug (see legend); only minimal and maximal concentrations are shown for the first strain of each species. Apart from colistin, the same concentration ranges were used for all *E. coli* and *K. pneumoniae* MDR strains. One of two replicates is shown. **b**, Drug synergies against the same MDR strains in the *G. mellonella* infection model. Larvae were infected by *E. coli* and *K. pneumoniae* MDR isolates (10^6 and 10^4 CFUs, respectively) and left untreated, treated with single drugs or with the drug combination. The percentage of surviving larvae was monitored at indicated intervals after infection. $n = 10$ larvae per treatment. The averages of four biological replicates are plotted; error bars depict s.d.



Extended Data Fig. 12 | Mode of action for the vanillin–spectinomycin synergy. **a**, The spectinomycin MIC decreases upon addition of $100 \mu\text{g ml}^{-1}$ vanillin in the wild-type *E. coli* BW25113, as well as in *E. coli* single-gene knockouts of members of the AcrAB–TolC efflux pump or its MarA regulator. Thus, the vanillin–spectinomycin synergy is independent of the effect of vanillin on AcrAB–TolC (Fig. 3). **b**, Synergy is specific to vanillin–spectinomycin, as spectinomycin is antagonized by $500 \mu\text{g ml}^{-1}$ of the vanillin-related compound aspirin, thereby increasing the MIC by approximately threefold. **c**, Profiling the vanillin–spectinomycin combination in the *E. coli* BW Keio collection²⁶ to deconvolute its mode of action. Violin plots of the drug–drug interaction scores (ε) of all mutants ($n = 9,216$; Methods) are presented for the vanillin–spectinomycin combination (synergy) and, as control, for the combination of vanillin with another aminoglycoside amikacin (antagonism). The interaction scores of the two *mdfA* deletion clones present in the Keio library are indicated by red dots. The vanillin–spectinomycin synergy is lost in the absence of *mdfA* but the vanillin–amikacin antagonism remains unaffected, which indicates that the vanillin–spectinomycin synergy depends specifically on MdfA. **d**, Deletion of *mdfA* leads to an increased spectinomycin MIC and abolishes the synergy with vanillin, independent

of the presence or absence of AcrAB–TolC. Mild overexpression of *mdfA* from a plasmid (pmdfA, Methods) further enhances the synergy with vanillin, decreasing the spectinomycin MIC by about twofold (compared to the MIC of the combination in the wild type). **e**, Overexpression of *mdfA* leads to increased spectinomycin sensitivity, even though the MIC does not change. The growth of *E. coli* BW25113 carrying a plasmid with *mdfA* cloned in it (pmdfA; no inducer, mild overexpression) or the empty vector (BW) was measured (OD_{595 nm} after 8 h) over twofold serial dilutions of spectinomycin and normalized to the no-drug growth of the corresponding strain (white and black dots represent the average of $n = 3$ independent biological replicates, error bars represent s.d.). The spectinomycin dose response was computed using a logistic fit of the averaged data points (MICs are calculated by fitting individual replicates, and then averaging). Fitted curves are represented by full and dashed lines for pmdfA and *E. coli* BW25113, respectively. **f**, Vanillin leads to accumulation of spectinomycin in the cell in an *mdfA*-dependent manner. Intracellular spectinomycin is measured with the tritiated compound (Methods). Bar plots and error bars in **a**, **b**, **d**, **f** represent the average and s.d., respectively, across n independent biological replicates.

The purinergic receptor P2RX7 directs metabolic fitness of long-lived memory CD8⁺ T cells

Henrique Borges da Silva^{1,2}, Lalit K. Beura^{1,3}, Haiguang Wang^{1,2}, Eric A. Hanse^{1,2}, Reshma Gore⁴, Milcah C. Scott^{1,3}, Daniel A. Walsh^{1,2}, Katharine E. Block^{1,2}, Raissa Fonseca^{1,3}, Yan Yan^{1,2}, Keli L. Hippen^{1,5}, Bruce R. Blazar^{1,5}, David Masopust^{1,3}, Ameeta Kelekar^{1,2}, Lucy Vulchanova⁴, Kristin A. Hogquist^{1,2} & Stephen C. Jameson^{1,2*}

Extracellular ATP (eATP) is an ancient ‘danger signal’ used by eukaryotes to detect cellular damage¹. In mice and humans, the release of eATP during inflammation or injury stimulates both innate immune activation and chronic pain through the purinergic receptor P2RX7^{2–4}. It is unclear, however, whether this pathway influences the generation of immunological memory, a hallmark of the adaptive immune system that constitutes the basis of vaccines and protective immunity against re-infection^{5,6}. Here we show that P2RX7 is required for the establishment, maintenance and functionality of long-lived central and tissue-resident memory CD8⁺ T cell populations in mice. By contrast, P2RX7 is not required for the generation of short-lived effector CD8⁺ T cells. Mechanistically, P2RX7 promotes mitochondrial homeostasis and metabolic function in differentiating memory CD8⁺ T cells, at least in part by inducing AMP-activated protein kinase. Pharmacological inhibitors of P2RX7 provoked dysregulated metabolism and differentiation of activated mouse and human CD8⁺ T cells *in vitro*, and transient P2RX7 blockade *in vivo* ameliorated neuropathic pain but also compromised production of CD8⁺ memory T cells. These findings show that activation of P2RX7 by eATP provides a common currency that both alerts the nervous and immune system to tissue damage, and promotes the metabolic fitness and survival of the most durable and functionally relevant memory CD8⁺ T cell populations.

P2RX7 is unique among the P2RX family in its activation by high concentrations of eATP (such as those released by dying cells)^{1,7}. Activation of P2RX7 induces ion transport (including Ca²⁺ influx and K⁺ efflux), but can also cause cell death by opening non-specific membrane pores^{2,4,8}. The results of gene ablation and pharmacological blockade of P2RX7 have suggested that it supports the activation and differentiation of certain subsets of effector CD4⁺ T cells, but induces the death of others^{7–10}. The role of P2RX7 in generating long-lived T cell memory is unclear. We evaluated the response of co-adoptively transferred wild-type and *P2rx7*-deficient (*P2rx7*^{−/−}) P14 T cell receptor (TCR) transgenic CD8⁺ T cells (P14 cells) following infection with acute lymphocytic choriomeningitis virus (LCMV Armstrong), and found a progressive defect in maintenance of *P2rx7*^{−/−} P14 cells in lymphoid tissues, despite normal expansion during the effector phase of the response (Fig. 1a–c). This resulted from a profound defect in the establishment of *P2rx7*^{−/−} CD62L⁺ central memory (T_{CM}) cells, while production of effector and CD62L[−] effector memory (T_{EM}) cells was much less affected (Fig. 1b, c). The skewing against *P2rx7*^{−/−} T_{CM} cells may be an underestimate, as P2RX7 signalling during wild-type cell isolation can provoke shedding of CD62L^{7,8}. Further separation of memory subsets extended these findings, confirming that T_{CM} cells were most severely compromised by P2RX7 deficiency (Extended Data Fig. 1d, e). Long-term CD8⁺ T cell memory is mediated by T_{CM} cells in lymphoid tissues but by tissue-resident memory (T_{RM}) cells in non-lymphoid sites⁶. *P2rx7*^{−/−} P14 cells showed defective T_{RM} cell generation, especially of the well-characterized CD69^{high}/CD103^{high}

T_{RM} population, in various non-lymphoid sites (Fig. 1d, Extended Data Fig. 1a–c). P2RX7 deficiency compromised T_{CM} and T_{RM} cell generation by polyclonal *P2rx7*^{−/−} CD8⁺ T cells and in response to other acute viral infections (Extended Data Fig. 1f, g), suggesting that these results are generalizable. Furthermore, *P2rx7*^{−/−} P14 cells showed an impaired response to chronic LCMV infection, especially generation of CXCR5⁺ cells (a subset of ‘exhausted’ CD8⁺ T cells that retain functionality)^{11–13} (Fig. 1e, f, Extended Data Fig. 1h). Cell-surface expression of P2RX7 increases following CD8⁺ T cell activation and is highest on T_{CM} and CD69^{high}/CD103^{high} T_{RM} cells (Extended Data Fig. 2a–f), corresponding with the effect of P2RX7 deficiency on these subsets. Similarly, reactivity to BzATP (an eATP analogue) correlated with P2RX7 expression on memory CD8⁺ T cell subsets, indicating that P2RX7 mediates the sensitivity of memory CD8 T cells to normal eATP (Extended Data Fig. 2g, h). Hence, the eATP sensor P2RX7 is essential for establishment of the most durable memory CD8⁺ T cell populations in lymphoid and many non-lymphoid tissues.

We next investigated the role of P2RX7 in early differentiation of memory CD8⁺ T cells. Although *P2rx7*^{−/−} P14 cells initially formed both short-lived effector cells (SLECs) and memory precursor effector cells (MPECs), this was followed by a selective decline in MPECs (Fig. 2a) whereas SLEC maintenance was largely unaffected (Fig. 2a, Extended Data Fig. 3a). RNA sequencing (RNA-seq) analysis revealed minimal changes in gene expression by *P2rx7*^{−/−} MPECs (Extended Data Fig. 3b), prompting us to evaluate other potential mechanisms of MPEC decline. Cellular metabolism is reprogrammed during the transition from activated to memory CD8⁺ T cells¹⁴, resulting in enhanced oxidative phosphorylation (OXPHOS), fatty acid oxidation and mitochondrial maintenance^{15,16}. Previous studies have suggested that P2RX7 regulates metabolic processes, although there is evidence for both beneficial and detrimental effects on cell viability^{7,17}. As early as eight days post-infection, *P2rx7*^{−/−} MPECs exhibited lower mitochondrial mass, mitochondrial membrane potential and glucose uptake, while SLECs were minimally affected (Fig. 2b, c, Extended Data Fig. 3c–g). Furthermore, in extracellular flux assays *P2rx7*^{−/−} MPECs (but not SLECs) exhibited a lower oxygen consumption rate (OCR) and reduced spare respiratory capacity—an indicator of the metabolic resilience of cells under stress, and a trait of CD8⁺ T_{CM} cells¹⁵ (Fig. 2d, e). These findings are consistent with proposals that calcium influx through P2RXs promotes mitochondrial activity in short-term activated T cells^{17,18}. Aerobic glycolysis (measured by extracellular acidification rate (ECAR)) was only slightly reduced in *P2rx7*^{−/−} MPECs (Extended Data Fig. 3h), such that the OXPHOS/aerobic glycolysis ratio (OCR/ECAR) of *P2rx7*^{−/−} MPECs resembled that of SLECs (Fig. 2f). The impact of P2RX7 deficiency persisted into the memory phase, *P2rx7*^{−/−} P14 cells displaying defective glucose uptake, mitochondrial maintenance and function, and fatty acid uptake. Most of these defects affected T_{CM} cells more substantially than T_{EM} cells (Fig. 2g–i, Extended Data Fig. 3i–m). Thus, our findings suggest that the eATP receptor P2RX7 is

¹Center for Immunology, University of Minnesota, Minneapolis, MN, USA. ²Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA. ³Department of Microbiology and Immunology, University of Minnesota, Minneapolis, MN, USA. ⁴Department of Neuroscience, University of Minnesota, Minneapolis, MN, USA. ⁵Department of Pediatrics, University of Minnesota, Minneapolis, MN, USA. *e-mail: james024@umn.edu

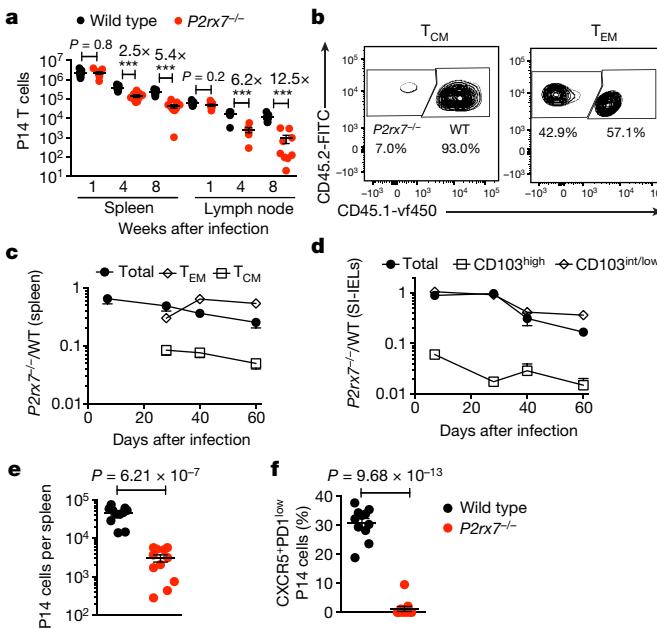


Fig. 1 | P2RX7 is required for the generation and maintenance of long-lived memory CD8⁺ T cells. **a–d**, Wild-type and *P2rx7*^{−/−} P14 cells were co-adoptively transferred and host mice were infected with LCMV Armstrong. **a**, Numbers of wild-type and *P2rx7*^{−/−} P14 cells in spleens and lymph nodes, 1, 4 and 8 weeks after infection. Fold-difference between wild-type and *P2rx7*^{−/−} P14 cell numbers is indicated. **b**, Frequencies of splenic wild-type (WT) and *P2rx7*^{−/−} P14 T_{CM} and T_{EM} cells 8 weeks after infection (representative of $n = 5$). **c**, **d**, Ratio of *P2rx7*^{−/−} P14 cells to wild-type P14 cells in T cell subsets in spleen (**c**) or small intestine intraepithelial lymphocytes (SI-IELs) (**d**). **e**, **f**, After co-adoptive transfer of wild-type and *P2rx7*^{−/−} P14 cells, mice were infected with LCMV-Cl13, and spleens analysed 2–3 weeks later for donor cell numbers (**e**) and percentage of CXCR5⁺PD1^{low} cells (**f**). All data shown as mean \pm s.e.m. **a–d**, Three independent experiments, $n = 4$ –9 total. **e**, **f**, Two independent experiments, $n = 11$ total. **a**, **e**, Two-tailed Student's *t*-test; **f**, two-tailed Mann–Whitney test; $^{**}P \leq 0.01$, $^{***}P \leq 0.001$.

critical for mitochondrial homeostasis and normal metabolic function in differentiating memory CD8⁺ T cells.

Metabolic defects might be predicted to compromise T cell survival. Indeed, basal proliferation of P2RX7-deficient memory P14 cells was normal, but they exhibited increased cell death (Extended Data Fig. 4a–c). Correspondingly, *P2rx7*^{−/−} MPECs and T_{CM} cells exhibited slightly reduced protein expression of the anti-apoptotic factor BCL2 and the transcription factors TCF1 and EOMES (associated with T_{CM} differentiation¹⁹), whereas SLEC and T_{EM} cell populations were much less affected (Extended Data Fig. 4d–f). Similar outcomes resulted when wild-type and *P2rx7*^{−/−} P14 cells were adoptively transferred separately (data not shown). Survival of T_{CM} cells is supported by interleukin (IL)-7 and IL-15⁵, but although *P2rx7*^{−/−} P14 cells displayed reduced cell surface expression of the IL-7 receptor IL-7R α , they showed normal dependence on IL-15 for homeostasis (Extended Data Fig. 4g–i). Furthermore, *P2rx7*^{−/−} P14 cells outcompeted wild-type cells for lymphopenia-induced proliferation (Extended Data Fig. 4j, k)—a homeostatic response to IL-7 and IL-15 that does not involve strong TCR stimulation²⁰. Together, these data suggest that impaired cytokine sensitivity is not the basis of defective *P2rx7*^{−/−} CD8⁺ T_{CM} cell maintenance.

To explore how P2RX7 selectively controls the metabolism of differentiating CD8⁺ memory T cells, we used in vitro assays in which activated CD8⁺ T cells cultured with IL-2 or IL-15 acquire effector- or memory-like properties, respectively^{15,21}. Wild-type and *P2rx7*^{−/−} P14 cells responded similarly to IL-2, but in IL-15 cultures *P2rx7*^{−/−} cells progressively declined in viability (Fig. 3a), analogous to defective memory survival in vivo (Extended Data Fig. 4c). Furthermore, after 72 h

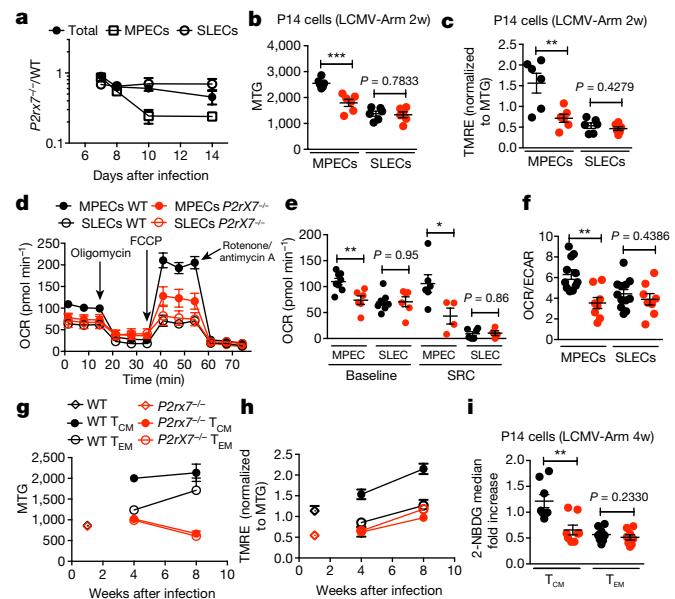


Fig. 2 | P2RX7 regulates mitochondrial homeostasis in memory CD8⁺ T cells. **a**, *P2rx7*^{−/−}/wild-type P14 cell ratios among MPECs and SLECs. **b**, **c**, Median fluorescence of Mitotracker Green (MTG; **b**) and tetramethylrhodamine (TMRE; normalized to MTG; **c**) in MPECs and SLECs. LCMV-Arm 2w, two weeks after infection with LCMV Armstrong. **d–f**, Sorted wild-type and *P2rx7*^{−/−} P14 MPECs and SLECs were analysed for metabolic function, indicated by OCR parameters (**d**, **e**) and OCR/ECAR ratios (**f**). **g–i**, Subsets of splenic wild-type and *P2rx7*^{−/−} P14 cells were stained for MTG (**g**) or TMRE (**h**) at the indicated times. **i**, Fold-change in median fluorescence of 2-(N-(7-nitrobenz-2-oxa-1,3-diazol-4-yl)amino)-2-deoxyglucose (2-NBDG), as a measure of glucose uptake, in indicated P14 memory cell subsets relative to naive CD8⁺ T cells. All data shown as mean \pm s.e.m. **a–c**, **g–i**, Three independent experiments, $n = 5$ –6 total per time point. **d–f**, Three independent experiments, cells pooled from five mice per experiment; $n = 4$ –6 total. **b**, **c**, **f–h**, Two-tailed Student's *t*-test; **e**, **i**, Two-tailed Mann–Whitney test; $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$.

of IL-15 culture, *P2rx7*^{−/−} P14 cells developed profound decreases in maximum OCR levels and spare respiratory capacity (SRC) (Fig. 3b, c) within the viable population (Extended Data Fig. 5a). IL-15-polarized *P2rx7*^{−/−} P14 cells also displayed defective aerobic glycolysis (Extended Data Fig. 5b) and decreased mitochondrial mass and membrane potential (Extended Data Fig. 5c, d). These results are unlikely to reflect altered cytokine sensitivity, as *P2rx7*^{−/−} and wild-type P14 cells showed similar pSTAT5 induction and re-expression of CD62L in IL-15 cultures²¹ (Extended Data Fig. 5e, f) and normal expansion (Extended Data Fig. 5g). IL-2-polarized *P2rx7*^{−/−} P14 cells showed some dysfunction (reduced mitochondrial mass and impaired aerobic glycolysis), suggesting that P2RX7 also alters effector metabolism, although *P2rx7*^{−/−} effector cell generation was normal in vivo (Fig. 2a). Hence, the ability of P2RX7 to control metabolism in nascent memory CD8⁺ T cells could be modelled in vitro.

Metabolic programming in T cells entails mitochondrial remodeling, and the development of fused mitochondria in memory CD8⁺ T cells correlates with more efficient OXPHOS²². Analysis of mitochondrial ultrastructure showed that IL-15-polarized *P2rx7*^{−/−} P14 cells fail to form the fused mitochondrial networks seen in wild-type cells, more closely resembling IL-2-cultured effector cells (Fig. 3d, Extended Data Fig. 6a). Furthermore, in vitro- and in vivo-activated *P2rx7*^{−/−} P14 cells exhibited increased proton leak during respiration at multiple stages of differentiation (Extended Data Fig. 6b–d), suggesting mitochondrial damage²². These data indicate that P2RX7 deficiency impairs the dynamic mitochondrial reorganization associated with memory CD8⁺ T cell differentiation. OPA1 drives mitochondrial fusion in T cells²², and OPA1 expression was reduced in IL-15-polarized *P2rx7*^{−/−}

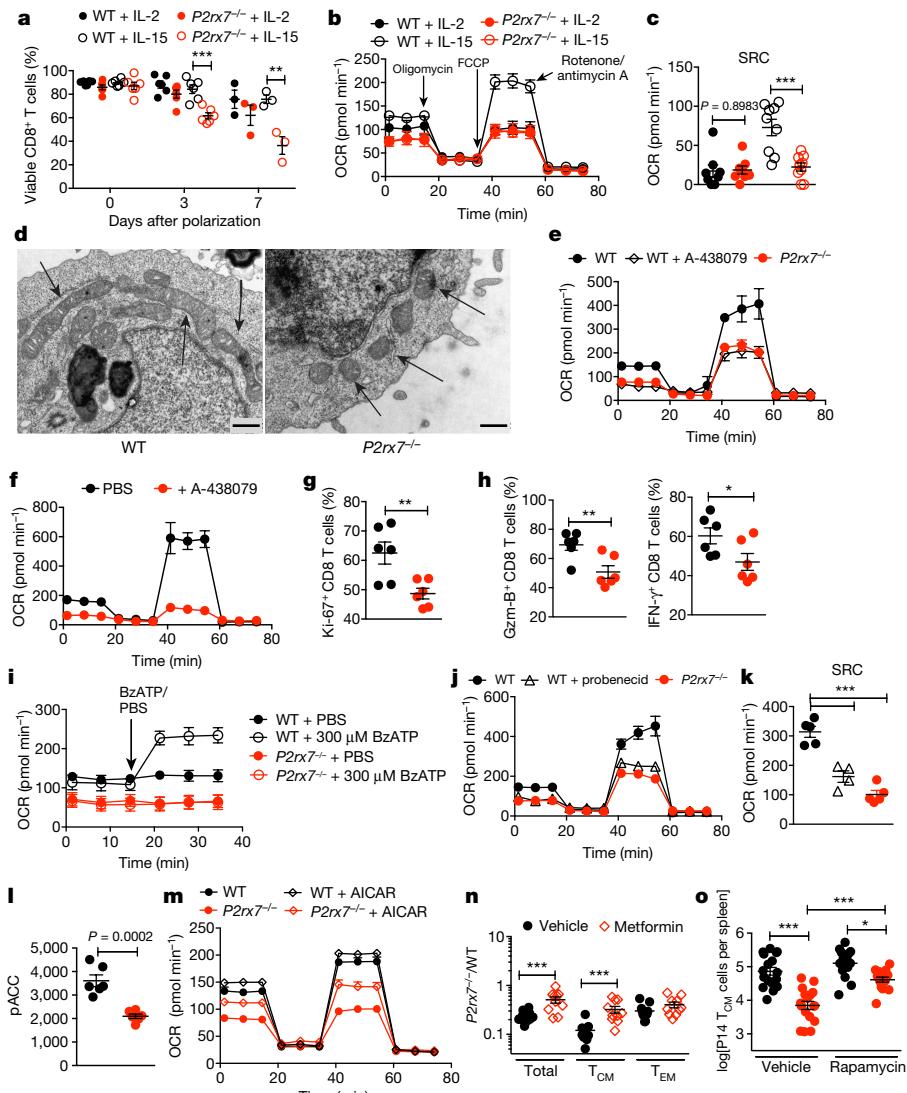


Fig. 3 | P2RX7 ablation leads to aberrant metabolism and depressed AMPK activation in CD8⁺ T cells. a-d, In vitro activated wild-type and P2rx7^{-/-} P14 cells were activated and then polarized with IL-2 or IL-15 (cells pooled from two mice per experiment). **a**, Cell viability during IL-2 or IL-15 cultures ($P \geq 0.2$ for all IL-2-polarized cells and IL-15-polarized cells at 24 h). **b**, **c**, OCR (**b**) and calculated SRC (**c**) for IL-2- and IL-15-polarized cells. **d**, Electron microscope images of mitochondria (arrows) in IL-15-polarized wild-type and P2rx7^{-/-} P14 cells (scale bars, 500 nm). **e-k**, Mouse (**e**, **i-k**) and human (**f-h**) CD8⁺ T cells were stimulated in vitro in the presence of A-438079 (**e-h**), BzATP (**i**), probenecid (**j**, **k**) or vehicle. Mouse cells activated as in **a**; human cells assayed 72 h after stimulation. OCR (**e**, **f**, **i-j**) and SRC (**k**) were measured and human cells

assayed for proliferation (Ki67) (**g**), granzyme B and IFN- γ (**h**). **l**, Median pACC in IL-15-polarized wild-type and P2rx7^{-/-} P14 cells. In **m**, cells were incubated for 6 h with or without AICAR before OCR measurement. **n**, **o**, Mice that received co-transferred wild-type and P2rx7^{-/-} P14 cells were infected with LCMV and treated with metformin (1–7 days post infection (d.p.i.); **n**) or rapamycin (4–8 d.p.i.; **o**). Ratios of P2rx7^{-/-} to wild-type cells for splenic memory subsets (**n**) and P14 T_{CM} cells (**o**). All panels, three independent experiments; $n = 3–6$ (**a**), 8–9 (**b**, **c**), 6 (**d**, **g**, **h**, **l**, **m**), 4–5 (**e**, **i**), 4 (**f**), 5–6 (**j**, **k**), 10–11 (**n**), 12–17 (**o**) total. All data shown as mean \pm s.e.m. **a**, **c**, **g**, **h**, **l**, **n**, Two-tailed Student's *t*-test; **k**, **o**, one-way ANOVA with Tukey's post-test; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$.

P14 cells (Extended Data Fig. 6e, f). However, OPA1 levels were proportional to total mitochondrial protein (Extended Data Fig. 6g, h) and it is unclear whether reduced OPA1 expression is a cause or consequence of dysregulated mitochondrial homeostasis in P2rx7^{-/-} CD8⁺ T cells.

It is possible that P2RX7 deficiency affected T cells before activation and differentiation. However, metabolism of naïve P2rx7^{-/-} CD8⁺ T cells was normal (Extended Data Fig. 7a–c). Furthermore, treatment with A-438079 (a highly specific P2RX7 inhibitor²³) during CD8⁺ T cell activation and IL-15 polarization resulted in dysregulated metabolism similar to that of P2rx7^{-/-} CD8⁺ T cells (Fig. 3e). These findings also provided an opportunity to extend our observations to human T cells. Treatment with A-438079 caused substantial loss of OXPHOS and aerobic glycolysis in activated human T cells (Fig. 3f,

Extended Data Fig. 7d), and a substantial decrease in their proliferation and cytokine production (Fig. 3g, h), but short-term survival was not altered (Extended Data Fig. 7e). Hence, while having some species-specific effects, blockade of P2RX7 compromises the metabolism of both human and mouse activated CD8⁺ T cells.

Our data suggested that eATP would promote CD8⁺ T cell metabolism and growth. Indeed, eATP blockade (using the ATP diphosphatase apyrase or the competitive inhibitor oATP) impaired expansion and OCR of IL-15-cultured P14 CD8⁺ T cells, while low concentrations of BzATP had the opposite effect (Fig. 3i, Extended Data Fig. 8a–e). Such doses of BzATP induced calcium mobilization but minimal pore formation (Extended Data Fig. 8f), consistent with studies suggesting that P2RX7-induced calcium flux enhances mitochondrial function^{7,10,17}. None of these compounds are P2RX7 specific, yet their effects on

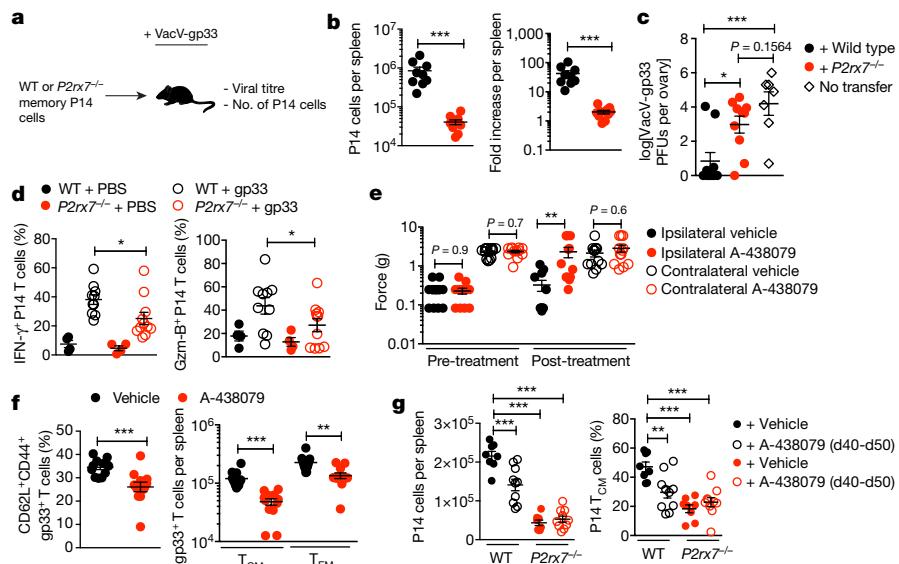


Fig. 4 | P2RX7 promotes memory CD8⁺ T cell function and P2RX7 blockade compromises CD8⁺ T cell memory maintenance. **a–c**, Wild-type or *P2rx7*^{−/−} memory P14 cells were independently transferred into naïve recipient mice, which were then infected with vaccinia virus encoding the LCMV epitope gp33 (VacV-gp33). **b**, Numbers of splenic wild-type or *P2rx7*^{−/−} P14 cells (left) and fold increase (right) seven days after infection. **c**, Titres of VacV-gp33 in ovaries seven days after infection (log₁₀-transformed). **d**, Wild-type and *P2rx7*^{−/−} P14 cells were independently transferred and recipient mice primed with LCMV. After more than 30 days, a transcervical challenge with gp33 or PBS was conducted and P14 cells in the female reproductive tract (FRT) were assayed 12 h later for IFN- γ production and granzyme B

(Gzm-B) expression. **e, f**, B6 mice subjected to spared nerve injury were subsequently infected with LCMV and treated with A-438079 or not treated. Mice were tested for pain sensitivity (**e**) and subsequent development of LCMV-specific (gp33⁺) splenic memory subsets (**f**). **g**, Mice were co-transferred and primed as in Fig. 1 and treated with A-438079 40–50 d.p.i. Left, number of P14 cells per spleen; right, per cent T_{CM} cells. **a–d**, Three independent experiments, $n = 9–10$ (**a, b**), 7–10 (**c**), 4–11 (**d**) total. **e–g**, Two independent experiments, $n = 12$ (**e, f**), 8–10 (**g**) total. All data shown as mean \pm s.e.m.; **b, e**, two-tailed Student's *t*-test; **e, f**, two-tailed Mann–Whitney *t*-test; **c, g**, one-way ANOVA with Tukey's post-test; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$.

CD8⁺ T cells were entirely P2RX7-dependent (Fig. 3i, Extended Data Fig. 8a–f). Furthermore, short-term P2RX7 inhibition blocked OCR in IL-15-cultured CD8⁺ T cells (Extended Data Fig. 8g). Hence, these studies suggest that physiological levels of eATP promote mitochondrial metabolism and growth of activated CD8⁺ T cells via a pathway that requires P2RX7.

An important issue was the source of eATP triggering P2RX7 signalling. While eATP is released from tissues during infection and injury^{1,7}, eATP is also exported by healthy activated T cells through Pannexin 1 (PANX1) channels, which may themselves be triggered by P2RX7 signalling^{18,24}. In this way, activated T cells could autonomously maintain P2RX7 signalling by both exporting and responding to eATP, perpetuating mitochondrial function through sustained calcium ion influx (and/or other functions of P2RX7). Both extrinsic and PANX1-derived eATP have been proposed to influence effector CD4⁺ T cell differentiation¹⁰. Notably, inclusion of the PANX1 inhibitor probenecid during in vitro generation of CD8⁺ T cell memory-like cells caused impaired OXPHOS and reduced SRC, similar to *P2rx7*^{−/−} CD8⁺ T cells (Fig. 3j, k), suggesting that PANX1 can regulate the same metabolic pathways. Consistent with a role for P2RX7 and PANX1 in ATP export, intracellular ATP concentrations were elevated and extracellular ATP concentrations were decreased in *P2rx7*^{−/−} effector CD8⁺ T cells, and PANX1 was critical for eATP (Extended Data Fig. 8h, i).

Elevated intracellular ATP concentrations could lead to impaired activation of AMP-activated protein kinase (AMPK), which controls adaptation to environmental stress²⁵, restrains mTOR activity and promotes the transition from effector to memory CD8⁺ T cells^{14,26}. Indeed, mTOR activity (indicated by pS6 levels) was elevated in memory *P2rx7*^{−/−} T cells (Extended Data Fig. 8j). More directly, IL-15-polarized *P2rx7*^{−/−} CD8⁺ T cells exhibited decreased phosphorylation of the AMPK target acetyl-CoA carboxylase (pACC; Fig. 3l). AMPK is also activated by intracellular calcium¹⁴, and stimulation with BzATP at doses that promote efficient calcium flux (Extended Data Fig. 8f)

increased the pACC/pS6 ratio in wild-type but not *P2rx7*^{−/−} CD8⁺ T cells (Extended Data Fig. 8k). Furthermore, in vitro treatment with AICAR (a pharmacological AMPK activator) largely corrected defective OCR and survival in *P2rx7*^{−/−} CD8⁺ T cells (Fig. 3m, Extended Data Fig. 8l). In vivo, activation of AMPK using metformin enhanced generation of *P2rx7*^{−/−} CD8⁺ T_{CM} and T_{RM} cells (while minimally affecting circulating T_{EM} cells), and partially or completely restored normal mitochondrial mass and membrane potential in *P2rx7*^{−/−} P14 T_{CM} cells (Fig. 3n, Extended Data Fig. 8m–o). Similarly, transient blockade of mTOR with rapamycin improved production of *P2rx7*^{−/−} T_{CM} cells (Fig. 3o, Extended Data Fig. 8p). Collectively, these data indicate that P2RX7 activates AMPK, which, perhaps via mTOR inhibition, supports the generation of long-lived memory CD8⁺ T cells.

In vitro cytotoxicity and granzyme B expression was normal in *P2rx7*^{−/−} P14 effector cells (data not shown), but the functionality of *P2rx7*^{−/−} memory CD8⁺ T cells has been unclear. Equivalent numbers of wild-type or *P2rx7*^{−/−} memory P14 cells were independently evaluated for response to vaccinia virus expressing gp33 (Fig. 4a), and *P2rx7*^{−/−} P14 cells showed reduced recall expansion (Fig. 4b) and viral control (Fig. 4c). Impaired control of LCMV clone 13 was also observed in polyclonal *P2rx7*^{−/−} mice (Extended Data Fig. 9a). *P2rx7*^{−/−} P14 cell recall responses to recombinant *Listeria* were also blunted, correlating with increased cell death rather than impaired proliferation (Extended Data Fig. 9b–f). Likewise, following local antigen challenge of female reproductive tract T_{RM} cells (using transcervical peptide stimulation²⁷), fewer *P2rx7*^{−/−} P14 T_{RM} cells produced IFN- γ or granzyme B, accompanied by reduced activation of 'bystander' CD8⁺ T cells and maturation of local dendritic cells (Fig. 4d, Extended Data Fig. 9g, h). These data demonstrate that P2RX7 is critical not only for maintenance of memory CD8⁺ T cells but also for their function.

P2RX7 is considered a promising pharmacological target for chronic pain treatment^{7,28}. Thus, we investigated whether P2RX7 blockade therapies designed to treat neuropathic pain would impede generation of CD8⁺ T cell memory, developing a combined model of spared nerve

injury with LCMV infection (Extended Data Fig. 9i–k). Transient in vivo treatment with A-438079 significantly attenuated nerve injury-induced hypersensitivity (Fig. 4e) and, in parallel, significantly decreased production of memory CD8⁺ T cells, especially T_{CM} cells, one month later (Fig. 4f). Furthermore, A-438079 treatment during the week following LCMV infection reduced subsequent generation of memory and MPEC (but not SLEC) P14 cells, resembling the defects of P2rx7^{-/-} CD8⁺ T cells (Extended Data Fig. 9l–n). A-438079 also impaired CD8⁺ T_{CM} and T_{RM} cell generation in LCMV infected BALB/c mice, which express a functionally distinct P2rx7 allele⁷ (Extended Data Fig. 9o). Notably, P2RX7 blockade caused loss of pre-existing memory CD8⁺ T cells, especially T_{CM} cells, suggesting that P2RX7 is required for maintenance of CD8⁺ T cell memory (Fig. 4g, Extended Data Fig. 9p). Hence, therapeutic P2RX7 inhibition may inadvertently compromise the development or maintenance of long-lived CD8⁺ T cell memory.

A paradigm shift in immunology came with understanding that detection of pathogen- and danger-associated molecular patterns are critical to spark immune reactivity^{29,30}. eATP is one of these triggers, representing a primordial mechanism for indicating tissue injury and inflammation¹, but the effect of this pathway on adaptive immune memory has been unclear. We show here that the eATP sensor P2RX7 has a hitherto unsuspected intrinsic role in supporting the generation of long-lived memory CD8⁺ T cells by driving their metabolic reprogramming and mitochondrial maintenance. Thus, eATP, produced by damaged tissue or exported by activated cells, not only triggers innate immune activation and inflammatory nociception but also promotes durable adaptive immunological memory (Extended Data Fig. 10).

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0282-0>.

Received: 21 September 2017; Accepted: 15 May 2018;

Published online 4 July 2018.

- Heil, M. & Land, W. G. Danger signals — damaged-self recognition across the tree of life. *Front. Plant Sci.* **5**, 578 (2014).
- Surprenant, A., Rassendren, F., Kawashima, E., North, R. A. & Buell, G. The cytolytic P2Z receptor for extracellular ATP identified as a P2X receptor (P2X7). *Science* **272**, 735–738 (1996).
- Chessell, I. P. et al. Disruption of the P2X7 purinoceptor gene abolishes chronic inflammatory and neuropathic pain. *Pain* **114**, 386–396 (2005).
- Mariathasan, S. et al. Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440**, 228–232 (2006).
- Williams, M. A. & Bevan, M. J. Effector and memory CTL differentiation. *Annu. Rev. Immunol.* **25**, 171–192 (2007).
- Schenkel, J. M. & Masopust, D. Tissue-resident memory T cells. *Immunity* **41**, 886–897 (2014).
- Di Virgilio, F., Dal Ben, D., Sarti, A. C., Giuliani, A. L. & Falzoni, S. The P2X7 receptor in infection and inflammation. *Immunity* **47**, 15–31 (2017).
- Rissiek, B., Haag, F., Boyer, O., Koch-Nolte, F. & Adriouch, S. P2X7 on mouse T cells: one channel, many functions. *Front. Immunol.* **6**, 204 (2015).
- Proietti, M. et al. ATP-gated ionotropic P2X7 receptor controls follicular T helper cell numbers in Peyer's patches to promote host-microbiota mutualism. *Immunity* **41**, 789–801 (2014).
- Trautmann, A. Extracellular ATP in the immune system: more than just a "danger signal". *Sci. Signal.* **2**, pe6 (2009).
- Utzschneider, D. T. et al. T cell factor 1-expressing memory-like CD8⁺ T cells sustain the immune response to chronic viral infections. *Immunity* **45**, 415–427 (2016).
- Im, S. J. et al. Defining CD8⁺ T cells that provide the proliferative burst after PD-1 therapy. *Nature* **537**, 417–421 (2016).
- He, R. et al. Follicular CXCR5-expressing CD8⁺ T cells curtail chronic viral infection. *Nature* **537**, 412–428 (2016).

- Pearce, E. L., Poffenberger, M. C., Chang, C. H. & Jones, R. G. Fueling immunity: insights into metabolism and lymphocyte function. *Science* **342**, 1242454 (2013).
- van der Windt, G. J. et al. Mitochondrial respiratory capacity is a critical regulator of CD8⁺ T cell memory development. *Immunity* **36**, 68–78 (2012).
- Buck, M. D., O'Sullivan, D. & Pearce, E. L. T cell metabolism drives immunity. *J. Exp. Med.* **212**, 1345–1360 (2015).
- Ledderose, C. et al. Mitochondrial dysfunction, depleted purinergic signaling, and defective T cell vigilance and immune defense. *J. Infect. Dis.* **213**, 456–464 (2016).
- Schenk, U. et al. Purinergic control of T cell activation by ATP released through pannexin-1 hemichannels. *Sci. Signal.* **1**, ra6 (2008).
- Chang, J. T., Wherry, E. J. & Goldrath, A. W. Molecular regulation of effector and memory T cell differentiation. *Nat. Immunol.* **15**, 1104–1115 (2014).
- Sprent, J. & Surh, C. D. Normal T cell homeostasis: the conversion of naive cells into memory-phenotype cells. *Nat. Immunol.* **12**, 478–484 (2011).
- Carrio, R., Bathe, O. F. & Malek, T. R. Initial antigen encounter programs CD8⁺ T cells competent to develop into memory cells that are activated in an antigen-free, IL-7- and IL-15-rich environment. *J. Immunol.* **172**, 7315–7323 (2004).
- Buck, M. D. et al. Mitochondrial dynamics controls T cell fate through metabolic programming. *Cell* **166**, 63–76 (2016).
- Donnelly-Roberts, D. L. & Jarvis, M. F. Discovery of P2X7 receptor-selective antagonists offers new insights into P2X7 receptor function and indicates a role in chronic pain states. *Br. J. Pharmacol.* **151**, 571–579 (2007).
- Saez, P. J. et al. ATP promotes the fast migration of dendritic cells through the activity of pannexin 1 channels and P2X7 receptors. *Sci. Signal.* **10**, eaah7107 (2017).
- Blagih, J. et al. The energy sensor AMPK regulates T cell metabolic adaptation and effector responses *in vivo*. *Immunity* **42**, 41–54 (2015).
- Pearce, E. L. et al. Enhancing CD8 T-cell memory by modulating fatty acid metabolism. *Nature* **460**, 103–107 (2009).
- Schenkel, J. M. et al. T cell memory. Resident memory CD8 T cells trigger protective innate and adaptive immune responses. *Science* **346**, 98–101 (2014).
- Bartlett, R., Stokes, L. & Sluyter, R. The P2X7 receptor channel: recent developments and the use of P2X7 antagonists in models of disease. *Pharmacol. Rev.* **66**, 638–675 (2014).
- Matzinger, P. Tolerance, danger, and the extended family. *Annu. Rev. Immunol.* **12**, 991–1045 (1994).
- Janeway, C. A., Jr & Medzhitov, R. Innate immune recognition. *Annu. Rev. Immunol.* **20**, 197–216 (2002).

Acknowledgements We thank the UMN Flow Cytometry Resource for cell sorting, C. Henzler (UMN Supercomputing Institute) for bioinformatics analysis, F. Zhou (UMN Characterization Facility) for transmission electron microscopy, M. Pierson for viral plaque assays, the NIH Tetramer Core for peptide/MHC tetramers, and A. Goldrath, S. Kaech, G. Shadel, R. Jones, E. Pearce, M. Jenkins, V. Vezys and members of the Jamequist laboratory and UMN Center for Immunology for discussions. The UMN Characterization Facility is a member of the NSF-funded Materials Research Facilities Network (<https://www.mrfn.org>) via the MRSEC program. This work was supported by NIH grants AI38903 and AI75168 (S.C.J.), CA157971 (A.K.), and MN Partnership Infrastructure Award MNPIF#16.09 (A.K.). H.B.D.S. was supported by a CNPq research fellowship from the Ministry of Science, Technology and Innovation of Brazil.

Reviewer information *Nature* thanks F. Grassi, J. Linden and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.C.J. and H.B.D.S. designed, analysed and interpreted the experiments. H.B.D.S., L.K.B., H.W., E.A.H., R.G., M.C.S., D.A.W., K.E.B., R.F. and Y.Y. performed experiments. R.G. performed spared nerve injury surgical procedures. E.A.H. and A.K. provided assistance with extracellular flux analysis. K.L.H., B.R.B., D.M., A.K., L.V. and K.A.H. contributed critical reagents and biological samples. S.C.J. and H.B.D.S. wrote the manuscript, with all authors contributing to editing the final text.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0282-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0282-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to S.C.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mice and infections. Six-to-eight-week old C57BL/6 (B6) and B6.SJL (expressing the CD45.1 allele) mice were purchased from Charles River (via the National Cancer Institute). *P2rx7*^{-/-}, *Rag2*^{-/-} and CX3CR1^{gfp/gfp} mice were obtained from Jackson Laboratories. LCMV-D^bGP33-specific TCR transgenic P14 mice were fully backcrossed to B6 and *P2rx7*^{-/-} mice, with introduction of CD45.1 and CD45.2 congenic markers for identification. Female mice were infected with LCMV Armstrong strain (2×10^5 PFU, intraperitoneally (i.p.)), LCMV clone 13 strain (2×10^6 PFU, intravenously (i.v.)) or VSV-OVA (1×10^6 PFU, i.v.). In some experiments, the LCMV viral titres in the kidneys of clone 13-infected mice were evaluated by standard plaque assay. In challenge experiments, mice were first infected with LCMV and, 6–8 weeks later, inoculated with *Listeria monocytogenes* (Lm)-GP33 (8×10^4 CFU). For vaccinia challenge experiments, wild-type and *P2rx7*^{-/-} memory P14 cells were sort-purified from CD45.1 congenic mice that had received wild-type or *P2rx7*^{-/-} P14 cells, respectively, and were infected with LCMV four weeks before being killed. Wild-type or *P2rx7*^{-/-} memory P14 cells (2×10^4) were separately transferred into naïve CD45.1 congenic recipient mice, which were then infected with VacV-gp33. After 7 days, mice were killed, spleens were removed for analysis of cell expansion, and ovaries were isolated for viral titre assessment by plaque assay. Viral titration was performed following a previously described protocol³¹. In brief, ovaries were homogenized in RPMI + 1% FBS containing penicillin and streptomycin with the aid of a handheld homogenizer, and serial dilutions of homogenates were plated on 143B cells (ATCC CRL-8304, mycoplasma status untested) for viral growth. After two days, plaques were enumerated after crystal violet staining. Transcervical challenges with GP33 peptide (50 µg per mouse) were performed as described³². Animals were maintained under specific-pathogen-free conditions at the University of Minnesota. All experimental procedures were approved by the institutional animal care and use committee at the University of Minnesota. No statistical methods were used to predetermine sample size. Animals were randomly allocated for infection and treatment groups. Investigators were not blinded unless otherwise indicated (see Spared nerve injury (SNI) model and measurement of mechanical sensitivity).

Administration of A-438079 to mice. A-438079 (Sigma-Aldrich) (in 0.5% DMSO in PBS) was administered to mice i.p. daily during the indicated treatment periods at a daily dose of 80 mg kg⁻¹ (which has been shown to inhibit nociceptive pain in murine models³³). Three different regimens of administration were used, in which A-438079 was given between days 1 and 7, between days 4 and 6 or between days 40 and 50 relative to LCMV infection. Control mice received daily vehicle injections (0.5% DMSO in PBS) during the same time periods.

Administration of metformin and rapamycin in mice. Metformin (Sigma-Aldrich) was administered to mice i.p. daily between days 1 and 7 after LCMV infection. The daily dose of metformin was 200 mg kg⁻¹, which has been shown to potentiate memory CD8⁺ T cell generation²⁶. Control mice received daily vehicle injections (0.1% DMSO in PBS) during the same period. Rapamycin (LC Laboratories) was administered to mice i.p. daily between days 4 and 8 after LCMV infection. The dose of rapamycin was 75 µg kg⁻¹ per day, as previously used during acute viral infection³⁴. Control mice received daily vehicle injections (0.01% DMSO in PBS) during the same period.

Flow cytometry. Lymphocytes were isolated from tissues including spleen, inguinal lymph nodes, cervical lymph nodes, blood, lung, gut intestinal epithelium, gut intestinal lamina propria, female reproductive tract, kidney and salivary glands as previously described^{35,36}. During isolation of lymphocytes from non-lymphoid tissues, 50 µg Treg-Protector (anti-ARTC2.2) nanobodies (BioLegend) were injected i.v. 15–30 min before mice were killed. Direct ex vivo staining and intracellular cytokine staining were performed as previously described^{36,37} with fluorochrome-conjugated antibodies (purchased from BD Biosciences, BioLegend, eBioscience, Cell Signaling Technology, Tonbo or Thermo Fisher Scientific). CXCR5 staining was performed as previously described¹². To detect LCMV-specific CD8⁺ T cell responses, tetramers were prepared as described³⁸. For discrimination of vascular-associated lymphocytes in non-lymphoid organs, in vivo i.v. injection of PE-conjugated CD8α antibody was performed as described³⁹. Among LCMV-specific CD8⁺ T cells, the following markers were used to distinguish these respective populations: T_{CM} (CD44⁺CD62L⁺), T_{EM} (CD44⁺CD62L⁻CD127⁺), T_{RM} (i.v. CD8α⁻CD69^{+/−}CD103^{high/int/low}), LLECs (CD44⁺CD62L⁻KLRG1⁺CX3CR1^{high}), MPECs (CD127⁺KLRG1⁻), and SLECs (CD127⁻KLRG1⁺). For detection of in vivo proliferation, cells were stained with Ki-67 using the Foxp3 kit for fixation and permeabilization. Alternatively, proliferation was assessed by BrdU incorporation as described⁴⁰. For survival assessment, cells were stained with Live/Dead (Tonbo Biosciences) and, when mentioned, with annexin V-FITC (eBioscience). For assessment of homeostatic proliferation, wild-type or *P2rx7*^{-/-} P14 cells were stained with CFSE and transferred into *Rag2*^{-/-} mice (2×10^6 cells per mouse) as described⁴¹. The numbers and percentages of CFSE^{low} cells were assessed within the next 3 weeks. After this period, mice were killed and the proliferation of spleen P14 cells was assessed. For measurement of

mitochondrial mass and membrane potential, cells were incubated with MTG (Thermo Fisher Scientific) and TMRE (Cell Signaling Technology) simultaneously for 15 min at 37 °C before ex vivo staining. For assessment of glucose uptake, cells were incubated with 50 µg ml⁻¹ 2-NBDG (Cayman Chemicals) for 2 h at 37 °C before ex vivo staining. For measurement of intracellular fatty acid levels, cells were incubated with 1 µg ml⁻¹ Bodipy^{493/503} (Thermo Fisher Scientific) for 20 min at 37 °C before ex vivo staining. For assessment of proliferation upon in vitro stimulation, CD8⁺ T cells were labelled with Cell Tracer Violet (CTV; Life Technologies) and after stimulation were stained ex vivo. For detection of intracellular factors such as BCL2, EOMES and TCF1, surface-stained cells were permeabilized, fixed and stained using the eBioscience Foxp3 staining kit, according to the manufacturer's instructions. For intracellular detection of pS6, ex vivo stained cells were fixed, permeabilized using the Phosflow Perm Buffer III (BD Biosciences) and stained with pS6 for 1 h at room temperature. For intracellular detection of pACC, cultured cells were fixed with paraformaldehyde 1%, permeabilized with 90% methanol and stained with pACC for 20 min at room temperature. Surface staining was performed after fixation to minimize effects on AMPK signalling. For assessment of calcium influx, cells were stained with surface markers, then incubated for 1 h at 37 °C with 1 µM Indo-1 (Thermo Fisher Scientific) and then washed, while for detection of large molecular weight pore formation, DAPI (Thermo Fisher Scientific; 1 µM) was added to the cell culture just before assay. In both cases, cells were analysed by kinetic flow cytometry, a baseline reading being determined for 1 min then BzATP (Sigma-Aldrich; 100 or 300 µM) added and the analysis continued for the remaining ~30 min. The ratios between bound Indo-1 and free Indo-1 were used as a measurement of calcium influx and uptake of DAPI was measured in the BV-421 channel. Flow cytometric analysis was performed on a LSR II or LSR Fortessa (BD Biosciences) and data were analysed using FlowJo software (Treestar).

Cell sorting. Cell sorting was performed on a FACS Aria III device (BD Biosciences). RNA-seq and extracellular flux analysis experiments were performed on KLRG1⁺CD127⁻ (SLEC) and KLRG1⁻CD127⁺ (MPEC) CD8⁺ T cells sorted from mice 8 or 14 days post-LCMV infection, respectively. The population purity after cell sorting was >95% in all experiments.

RNA-seq analysis. MPECs and SLECs were first homogenized eight days after LCMV infection using QIAshredder columns (Qiagen) and RNA was then extracted using an RNeasy kit (Qiagen) following the manufacturer's instructions. After quality control, total RNA samples were processed with the Illumina TotalPrep-96 RNA Amplification Kit for HighThroughput RNA Amplification for Array Analysis, and read in a HiSeq 2500 System (high output sequencing, paired end read, 125 bp).

Metabolic assays. OCR and ECAR were assessed using a 96-well XF Extracellular flux analyser, according to the manufacturer's instructions (Seahorse Bioscience). SRC, OCR/ECAR ratios and proton leak were defined as previously described^{15,22}. In some extracellular flux assays, BzATP (300 µM), oATP (EMD Millipore; 50 µM) or apyrase (New England Biolabs; 10 U ml⁻¹) were added either at the port or 1 h before analysis. In other experiments, A-438079 (25 µM) was added 6 h prior to analysis. Intracellular ATP concentrations were measured by using the ATP determination kit (Life Technologies). To assess the levels of extracellular ATP, wild-type or *P2rx7*^{-/-} P14 cells were activated in vitro and polarized with IL-2 or IL-15 as described above, except that the cytokine polarization cultures were performed in Transwells (Corning). After 24 h of polarization, the transwells were transferred to new wells and the plates were centrifuged at 50 r.p.m. for 1 min. The supernatant below the transwells was quantified for ATP concentration as described above. In all transwells, a combination of the plasma membrane ATPase inhibitor ebselein (30 µM; Cayman Chemical) and the ecto-ATPase inhibitor ARL 67156 trisodium salt hydrate (100 µM; Sigma-Aldrich) was added 1 h before evaluation, to inhibit ATPase activity⁴². In some samples, a Pannexin 1 inhibitor (¹⁰Panx, Tocris) was added 20 h before assessment.

Transmission electron microscopy. Cells were fixed in 2.5% glutaraldehyde in 100 mM sodium cacodylate, followed by post-fixation in 1% osmium tetroxide. After extensive washing, samples were stained in 1% uranyl acetate for 30 min in the dark, then washed again. Samples were dehydrated in ethanol and embedded in eponate resin. Cells were imaged using a JEOL 1200 EXII transmission electron microscope equipped with a SIS MegaView III high resolution CCD camera (1,376 × 1,032-pixel format, 12 bit). Mitochondrial areas were calculated using ImageJ software (NIH).

Protein quantification. Western blotting of total protein lysates was performed as previously described²². The following antibodies were used: Opa1 (BD Biosciences) and β-actin (Cell Signaling Technologies). Primary incubations were followed by incubation with secondary HRP-conjugated antibodies (Santa Cruz Biotechnology). Blots were revealed using the Biomax MR film (Kodak).

To quantify the total mitochondrial protein amounts, mitochondria were isolated from experimental P14 cells by using the Qproteome Mitochondria Isolation Kit (QIAgen), following the manufacturer's instructions. The mitochondria protein extracts were quantified using the Pierce BCA protein kit (Thermo Fisher).

In vitro culture experiments. P14 splenocytes from naive mice were stimulated for 72 h with gp33 peptide (1 μ M, KAVYNFATM, New England Peptide) and IL-2 (10 ng ml $^{-1}$). When indicated, probenecid (100 μ M, Sigma-Aldrich) or A-438079 (25 μ M) was added to the cultures (48 h before analysis). Where indicated, P14 cells were isolated with the mouse CD8 $^{+}$ T cell isolation kit (Miltenyi Biotech) 72 h after initial activation, and incubated for an additional 72 h or 7 days with IL-2 or IL-15 to induce generation of effector-like and memory-like populations, as previously described²¹. Cell numbers were assessed by Neubauer chamber counting. Survival was analysed by Trypan blue staining during microscope counting, and confirmed by Live-Dead (Tonbo Biosciences) staining using flow cytometry. For all experiments, complete RPMI medium (RPMI 1640 supplemented with 10% FBS, 100 U ml $^{-1}$ penicillin/streptomycin, 2 mM L-glutamine) was used. In some experiments, IL-2- or IL-15-cultured P14 cells were cultured in the presence of 100 μ M-1 mM BzATP for 72 h and cell numbers and survival were measured as indicated. In other experiments, IL-2- or IL-15-cultured P14 cells were cultured in the presence of either 50–100 μ M oATP or 10 U ml $^{-1}$ apyrase for 72 h, and cell numbers and survival were measured as indicated. In some experiments, IL-2- or IL-15-cultured P14 cells were cultured in the presence of 500 μ M AICAR (Sigma-Aldrich) for 72 h, and cell numbers and survival were measured as indicated. For extracellular flux analysis, AICAR was added 6 h before assay. Survival of wild-type and *P2rx7* $^{-/-}$ cells in extracellular flux assay medium was assessed by the WST-1 viability assay according to the manufacturer's instructions (Sigma-Aldrich).

Human CD8 $^{+}$ T cell studies. Resting CD8 $^{+}$ T cells were isolated from Ficoll-purified PBMCs isolated from leukapheresis products (Memorial Blood Center, St. Paul, MN) in a two-step process by first depleting CD25 $^{+}$ cells using (cGMP)-grade anti-CD25 microbeads (Miltenyi Biotech) on an AutoMACS (Miltenyi Biotech). CD8 $^{+}$ cells were then purified from the CD25 $^{-}$ fraction using negative selection (CD8 $^{+}$ T cell isolation kit, Miltenyi Biotech). Purified cells were stimulated for 72 h with α CD3/ α CD28 beads (InVitrogen) and IL-2 (300 U ml $^{-1}$). After 20 h, A-438079 (25 μ M) or PBS was added to the cultures. Cell numbers were analysed by Neubauer chamber counting. Survival was analysed by Trypan blue staining during microscope counting and confirmed by Live-Dead (Tonbo Biosciences) staining using flow cytometry.

Spared nerve injury (SNI) model and measurement of mechanical sensitivity. The SNI model produces substantial and prolonged changes in mechanical sensitivity and cold responsiveness that closely mimic the cutaneous hypersensitivity associated with clinical neuropathic pain⁴³. SNI surgeries were performed under isoflurane anaesthesia as described⁴⁴. Fourteen days after SNI the mice were infected with LCMV and treated with A-438079 (80 mg kg $^{-1}$) or vehicle on days 1–7 post-infection. For assessment of mechanical sensitivity, mechanical withdrawal thresholds were tested using von Frey filaments. Mice were acclimated for at least 30 min in the testing environment within a plastic box on a raised metal mesh platform. A logarithmically increasing set of eight von Frey filaments (Stoelting), labelled from 0.07 to 6.0 g, were used. These were applied perpendicular to the ventral-lateral hindpaw surface with sufficient force to cause a slight bending of the filament. A rapid withdrawal of the paw away from the stimulus fibre within 4 s was deemed as a positive response. Using an up-down statistical

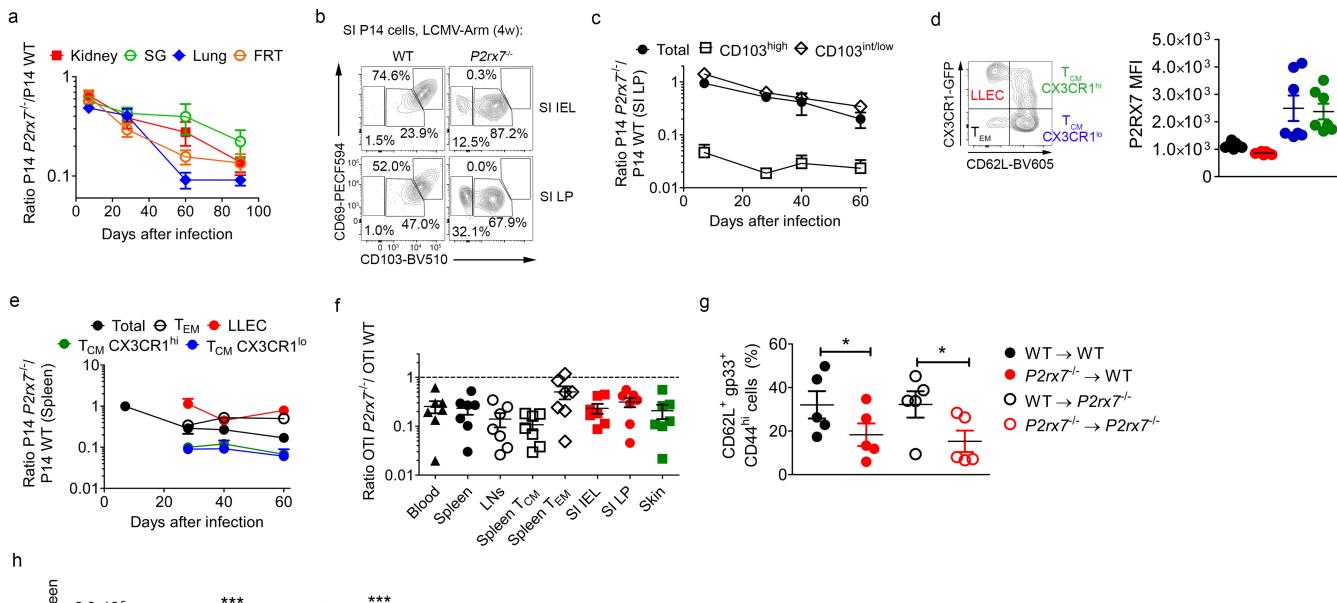
method⁴⁵, the 50% withdrawal threshold was calculated for each mouse and then averaged within the experimental groups. Mechanical withdrawal thresholds were measured at baseline, on day 14 post-SNI, and 6 h after the last drug treatment. As these measurements involved immediate investigator interpretation, investigator blinding to the group identities was used.

Statistical analysis. Data were subjected to the Kolmogorov-Smirnov test to assess Gaussian distribution. Statistical differences were calculated by using unpaired two-tailed Student's *t*-test or one-way ANOVA with Tukey's post-test where indicated. All experiments were analysed using Prism 5 (GraphPad Software). *P* values of $<0.05^*$, $<0.01^{**}$ or $<0.001^{***}$ indicate significant differences between groups. Exact *P* values for all experiments are provided in Supplementary Table 1.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

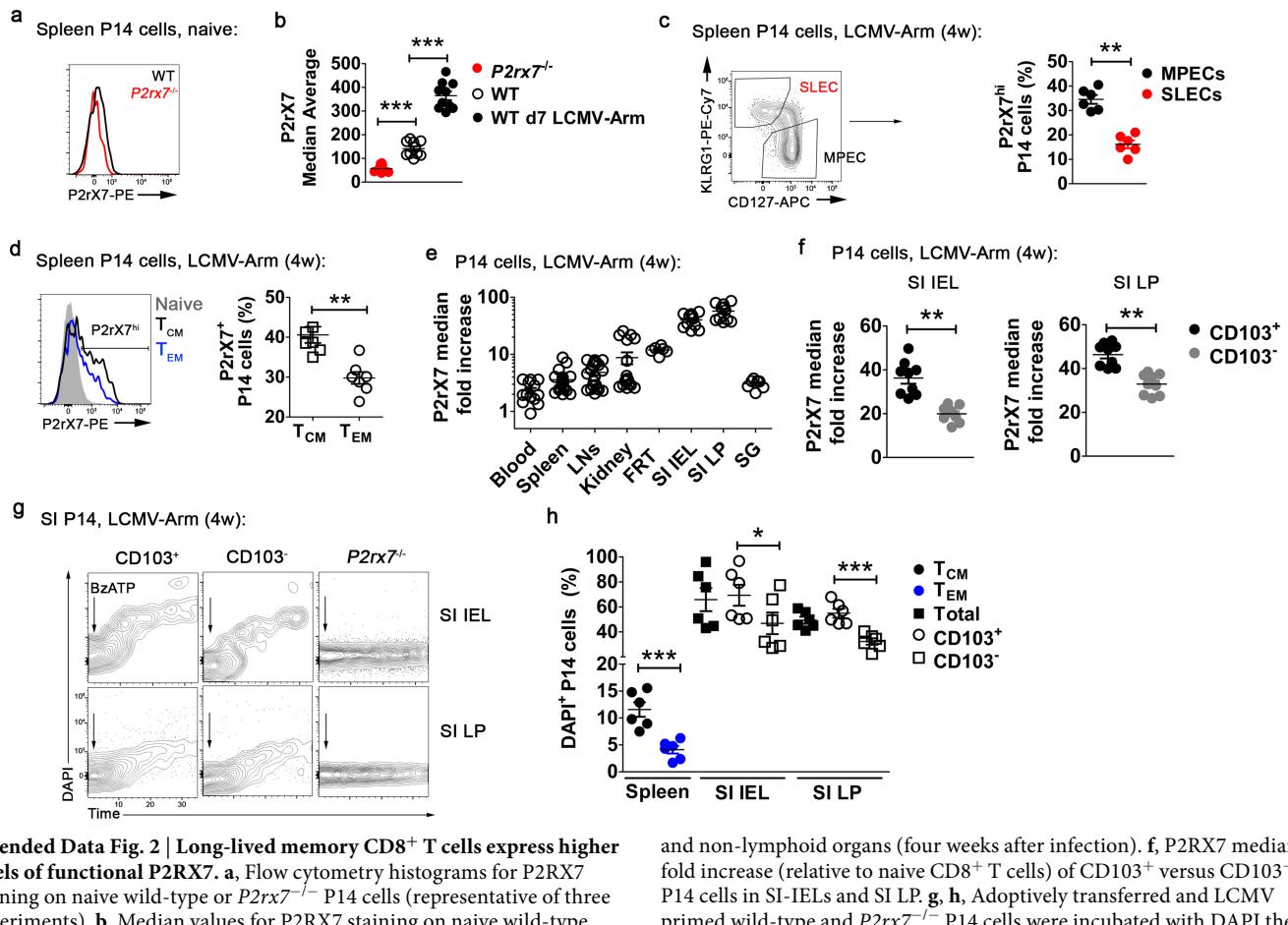
Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

31. Thompson, E. A., Beura, L. K., Nelson, C. E., Anderson, K. G. & Vezys, V. Shortened intervals during heterologous boosting preserve memory CD8 T cell function but compromise longevity. *J. Immunol.* **196**, 3054–3063 (2016).
32. Schenkel, J. M., Fraser, K. A., Vezys, V. & Masopust, D. Sensing and alarm function of resident memory CD8 $^{+}$ T cells. *Nat. Immunol.* **14**, 509–513 (2013).
33. McGaughy, S. et al. P2X $_7$ -related modulation of pathological nociception in rats. *Neuroscience* **146**, 1817–1828 (2007).
34. Araki, K. et al. mTOR regulates memory CD8 T-cell differentiation. *Nature* **460**, 108–112 (2009).
35. Steinert, E. M. et al. Quantifying memory CD8 T cells reveals regionalization of immunosurveillance. *Cell* **161**, 737–749 (2015).
36. Skon, C. N. et al. Transcriptional downregulation of S1pr1 is required for the establishment of resident memory CD8 $^{+}$ T cells. *Nat. Immunol.* **14**, 1285–1293 (2013).
37. Renkema, K. R. et al. IL-4 sensitivity shapes the peripheral CD8 $^{+}$ T cell pool and response to infection. *J. Exp. Med.* **213**, 1319–1329 (2016).
38. Daniels, M. A. & Jameson, S. C. Critical role for CD8 in T cell receptor binding and activation by peptide/major histocompatibility complex multimers. *J. Exp. Med.* **191**, 335–346 (2000).
39. Anderson, K. G. et al. Intravascular staining for discrimination of vascular and tissue leukocytes. *Nat. Protocols* **9**, 209–222 (2014).
40. Schenkel, J. M. et al. IL-15-independent maintenance of tissue-resident and boosted effector memory CD8 T cells. *J. Immunol.* **196**, 3920–3926 (2016).
41. Kieper, W. C. & Jameson, S. C. Homeostatic expansion and phenotypic conversion of naïve T cells in response to self peptide/MHC ligands. *Proc. Natl. Acad. Sci. USA* **96**, 13306–13311 (1999).
42. Praetorius, H. A. & Leipziger, J. ATP release from non-excitable cells. *Purinergic Signal.* **5**, 433–446 (2009).
43. Decosterd, I. & Woolf, C. J. Spared nerve injury: an animal model of persistent peripheral neuropathic pain. *Pain* **87**, 149–158 (2000).
44. Bourquin, A. F. et al. Assessment and analysis of mechanical allodynia-like behavior induced by spared nerve injury (SNI) in the mouse. *Pain* **122**, 14.e11–14 (2006).
45. Chaplan, S. R., Bach, F. W., Pogrel, J. W., Chung, J. M. & Yaksh, T. L. Quantitative assessment of tactile allodynia in the rat paw. *J. Neurosci. Methods* **53**, 55–63 (1994).



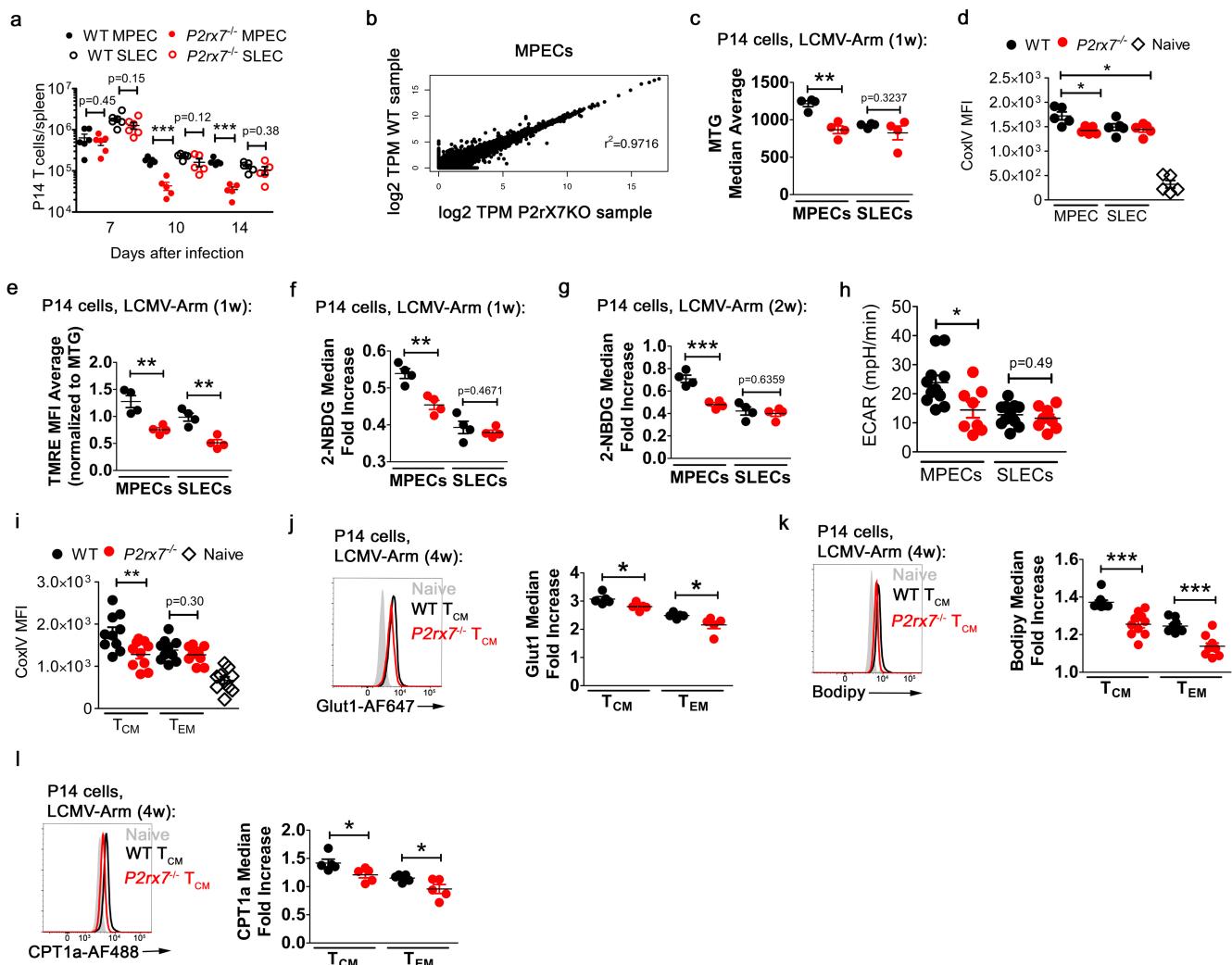
Extended Data Fig. 1 | P2RX7 is required for CD103^{high} T_{RM} and T_{CM} cell generation upon acute viral infection, and for establishment of Ag-specific CD8⁺ T cells upon chronic viral infection. **a–e**, Wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells were mixed 1:1 and co-adoptively transferred into B6.SJL mice that were subsequently infected with LCMV, and donor cells were identified as in Fig. 1. Data from 3–4 independent experiments, $n = 4$ FRT and $n = 5$ –7 other organs from all experiments. **a**, Ratios of *P2rx7*^{−/−} to wild-type P14 cells in different non-lymphoid organs over time. **b**, Flow cytometric plots showing CD69 and CD103 co-expression by wild-type and *P2rx7*^{−/−} P14 cells from SI-IELs and SI LP (small intestine lamina propria) 4 weeks post-infection (representative of three experiments). **c**, Ratios of *P2rx7*^{−/−} to wild-type P14 cells for CD103^{high}, CD103^{int} and CD103^{low} subsets among SI LP over time. **d**, **e**, We also evaluated the role of P2RX7 in the generation and maintenance of memory CD8⁺ T cell subsets based on CX3CR1 expression. **d**, Representative plot depicting the subpopulations studied (left) and P2RX7 median in these subsets (right). **e**, Ratio of *P2rx7*^{−/−} to

wild-type P14 CD8⁺ T cells in spleen, gated on indicated subsets. Data from 2–3 independent experiments, $n = 4$ –7 mice total. **f**, CD8⁺ T cells from wild-type and *P2rx7*^{−/−} OT-I TCR transgenic (OT-I) mice were mixed 1:1 and co-adoptively transferred into B6 mice subsequently infected with VSV-OVA (two independent experiments, $n = 7$ from all experiments). Ratio of wild-type to *P2rx7*^{−/−} OT-I cells in indicated tissues was determined 4 weeks after VSV-OVA infection. **g**, The indicated radiation bone marrow chimaeras were generated and infected with LCMV. Percentages of splenic D^b/gp33-tetramer binding (gp33⁺) CD8⁺ T_{CM} cells were determined 8 weeks post-infection with LCMV (data from two independent experiments, $n = 5$ from all experiments). **h**, Wild-type or *P2rx7*^{−/−} mice were infected with LCMV-Arm or LCMV-Cl13, and the numbers (left) and percentages of CXCR5⁺ PD1^{low} (right) gp33⁺ CD8⁺ T cells were evaluated 4 weeks after infection ($n = 6$ –13 from all experiments). **a, c–h**, Mean \pm s.e.m.; **g, h**, two-tailed Student's *t*-test, $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$.



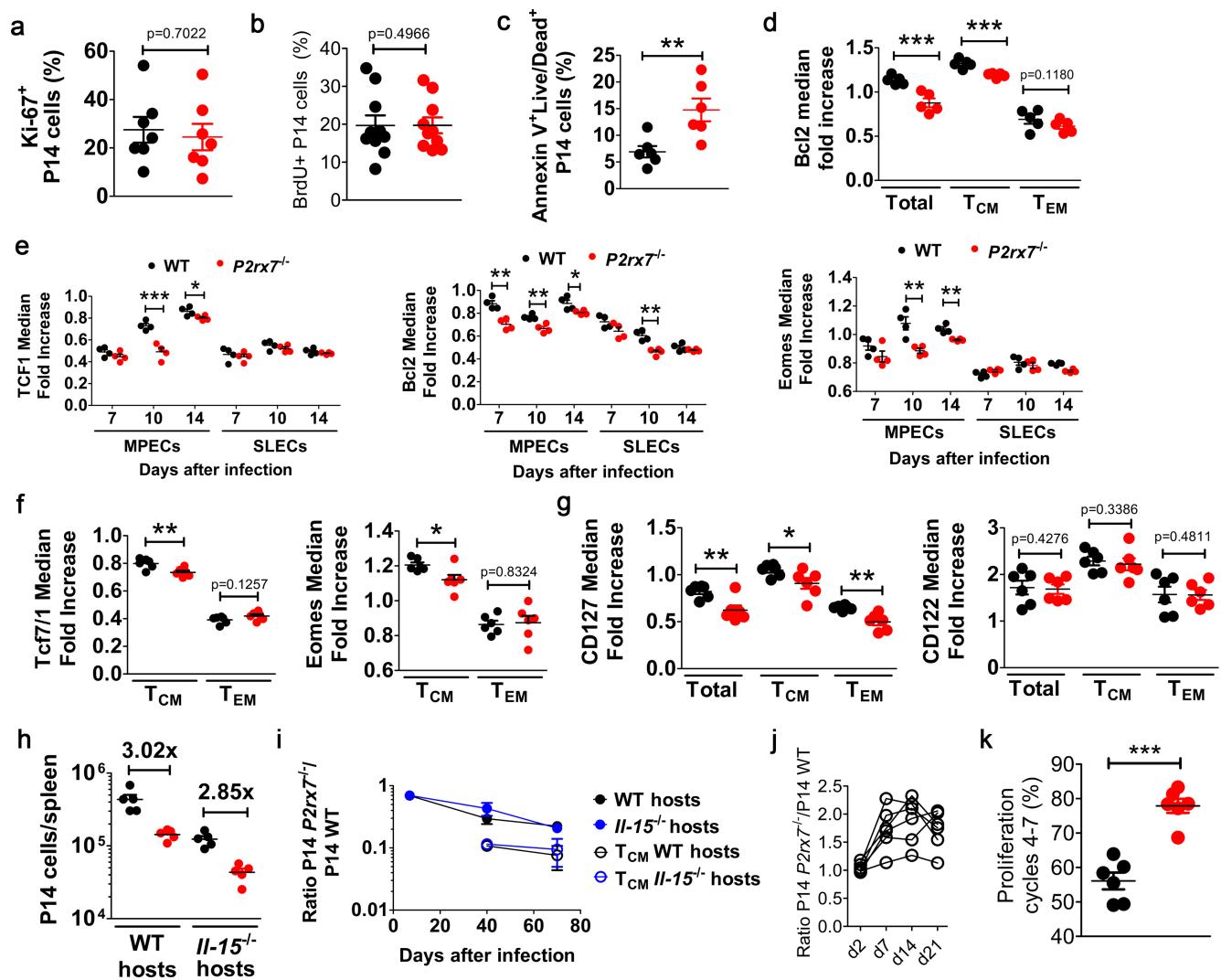
Extended Data Fig. 2 | Long-lived memory CD8⁺ T cells express higher levels of functional P2RX7. **a**, Flow cytometry histograms for P2RX7 staining on naive wild-type or *P2rx7*^{-/-} P14 cells (representative of three experiments). **b**, Median values for P2RX7 staining on naive wild-type and *P2rx7*^{-/-} P14 cells, and on wild-type P14 cells 7 d post-LCMV infection (data from three independent experiments, $n = 10$ total). **c-h**, Indicated subsets of adoptively transferred wild-type P14 CD8⁺ T cells from listed tissues were assayed for P2RX7 expression and functional response to the P2RX7 agonist BzATP following priming with LCMV for the indicated time (data from three independent experiments, $n = 6-21$ from all experiments). **c, d**, Percentage of P2RX7^{high} wild-type P14 cells in MPEC and SLEC (c) and T_{CM} and T_{EM} subsets (d) four weeks after infection. **e**, P2RX7 median fold increase (relative to expression in naive CD8⁺ T cells, showed as a dashed line) of P14 cells in blood, lymphoid

and non-lymphoid organs (four weeks after infection). **f**, P2RX7 median fold increase (relative to naive CD8⁺ T cells) of CD103⁺ versus CD103⁻ P14 cells in SI-IELs and SI LP. **g, h**, Adoptively transferred and LCMV primed wild-type and *P2rx7*^{-/-} P14 cells were incubated with DAPI then stimulated during flow cytometry with 300 μ M BzATP (which mediates P2RX7 pore opening and DAPI uptake). **g**, Flow cytometry plots of DAPI uptake by SI-IELs and SI LP wild-type and *P2rx7*^{-/-} P14 CD8⁺ T cells over 30 min (representative of six samples). **h**, Compiled data for percentage DAPI⁺ P14 cells (defined as the percentage above *P2rx7*^{-/-} DAPI levels 5 min after BzATP stimulation) in spleen, SI IEL and SI LP P14 cells ($n = 6$ from all experiments). **b-f, h**, Mean \pm s.e.m.; **b**, one-way ANOVA with Tukey's post-test; **c, d, f, h**, two-tailed Student's *t*-test; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$.



Extended Data Fig. 3 | P2RX7 is required for optimal metabolism of MPECs and T_{CM} cells. Wild-type and P2rx7^{-/-} P14 CD8⁺ T cells were mixed 1:1 and co-adoptively transferred into B6.SJL mice that were subsequently infected with LCMV. Donor cells were identified as in Fig. 1. **a**, Numbers of wild-type and P2rx7^{-/-} P14 MPECs and SLECs in spleens (three independent experiments, $n = 5$ –6 from all experiments). **b**, Gene expression profile from wild-type versus P2rx7^{-/-} P14 MPECs (sorted two weeks post-infection with LCMV). Gene expression in SLEC populations, sorted and analysed at the same time, also showed minimal differences between wild-type and P2rx7^{-/-} groups (data not shown). **c–g**, At day 8 (**c**, **e**, **f**) or 14 (**d**, **g**), splenic wild-type and P2rx7^{-/-} MPEC and SLEC subpopulations were stained for MTG (**c**), Cox-IV (**d**), TMRE (**e**)

or 2-NBDG uptake (**f**, **g**); two independent experiments, $n = 4$ –5 from all experiments. In **h**, the ECAR levels in MPECs and SLECs (from the experiments described in Fig. 2d–f) are shown (three independent experiments, $n = 8$ –12). **i–l**, Four weeks after priming with LCMV, wild-type and P2rx7^{-/-} P14 CD8⁺ T_{CM} and T_{EM} subsets were assessed for expression of Cox-IV (**i**), Glut1 (**j**) or CPT1a (**l**), or uptake of Bodipy (**k**). **j–l**, Left, representative flow cytometric plots for T_{CM} populations (relative to naive wild-type host CD8⁺ T cells); right, median difference in staining relative to naive wild-type host CD8⁺ T cells. **i–l**, Three independent experiments, $n = 5$ –10 from all experiments. **a**, **c–l**, Mean \pm s.e.m.; **a**, **c**, **e–l**, two-tailed Student's *t*-test; **d**, one-way ANOVA with Tukey's post-test; $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$.



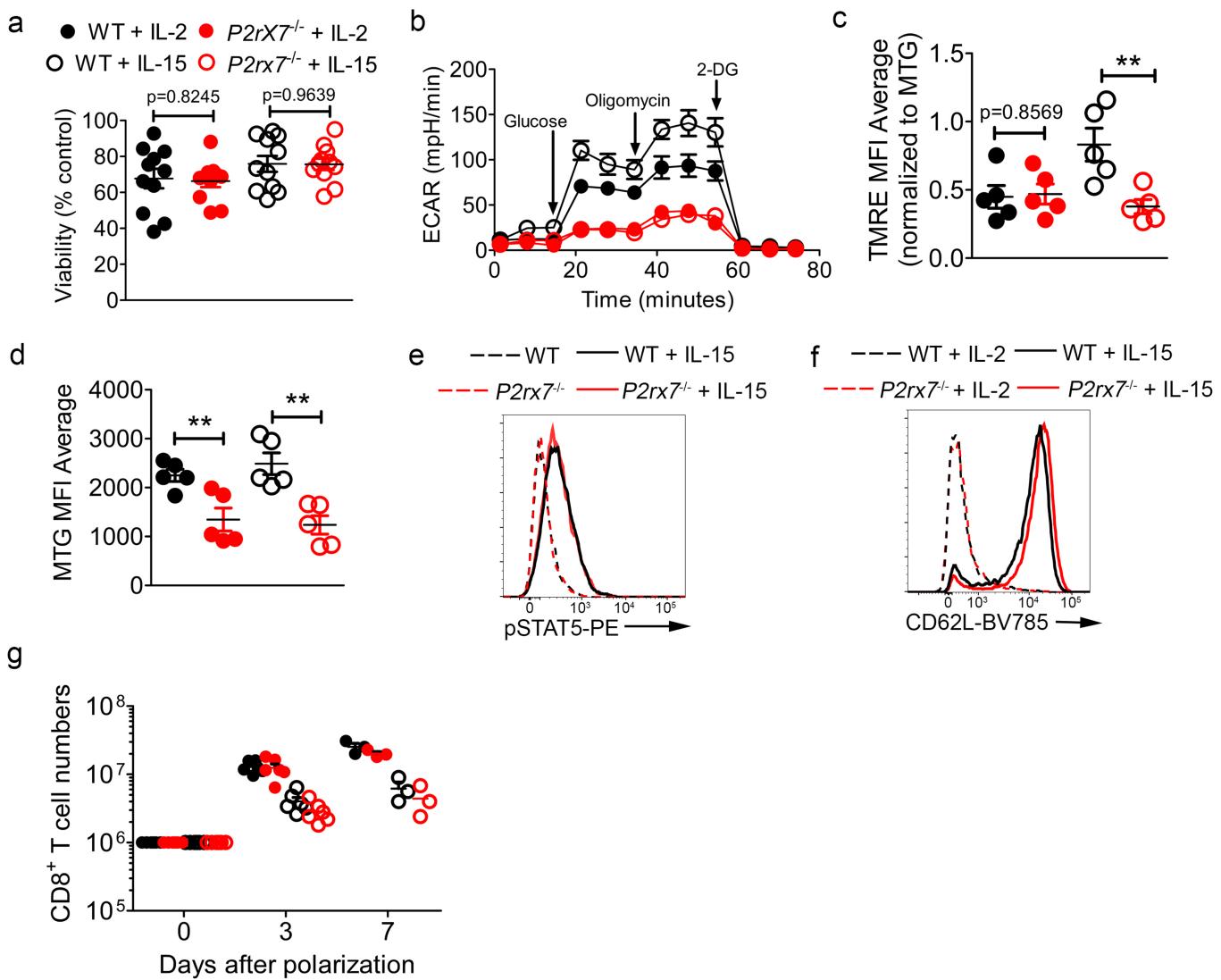
Extended Data Fig. 4 | P2RX7 signalling is required for survival rather than homeostatic proliferation of long-lived memory CD8⁺ T cells.
a–g, Wild-type (black) and *P2rx7*^{−/−} (red) P14 CD8⁺ T cells were mixed 1:1 and co-adoptively transferred into B6.SJL mice that were subsequently infected with LCMV, and donor cells were identified as in Fig. 1 (data from three independent experiments, $n = 4$ –10 from all experiments). **a–d**, Percentages of Ki-67⁺ cells (ex vivo staining; **a**), BrdU⁺ cells (**b**) and annexin V⁺ Live–Dead⁺ cells (**c**), and BCL2 median fold increase (relative to median values of naïve CD8⁺ T cells) for bulk, T_{CM} and T_{EM} subsets (**d**) was determined for wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells eight weeks after infection. **e**, Median fold increase (relative to naïve CD8⁺ T cells) in expression of TCF1 (left), BCL2 (centre) and EOMES (right) in wild-type and *P2rx7*^{−/−} P14 MPEC and SLEC CD8⁺ T cells at the indicated times. **f**, Expression of TCF1 (left) and EOMES (right) (shown as median expression relative to naïve CD8⁺ T cells) in splenic T_{EM} and T_{CM} subsets of wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells, four weeks after infection. **g**, CD127 (left) and CD122 (right) median expression (relative to naïve

CD8⁺ T cells) for splenic bulk cells and T_{CM} and T_{EM} subsets four weeks after infection. **h**, **i**, Wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells were mixed 1:1 and co-adoptively transferred into B6.SJL or IL-15^{−/−} mice, which were subsequently infected with LCMV (data from two independent experiments, $n = 4$ –5 from all experiments). **h**, Numbers of wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells in each host (four weeks after infection).

i, Ratio of total and T_{CM} *P2rx7*^{−/−} to wild-type P14 CD8⁺ T cells in spleens from wild-type and IL-15^{−/−} hosts at indicated times post-infection.

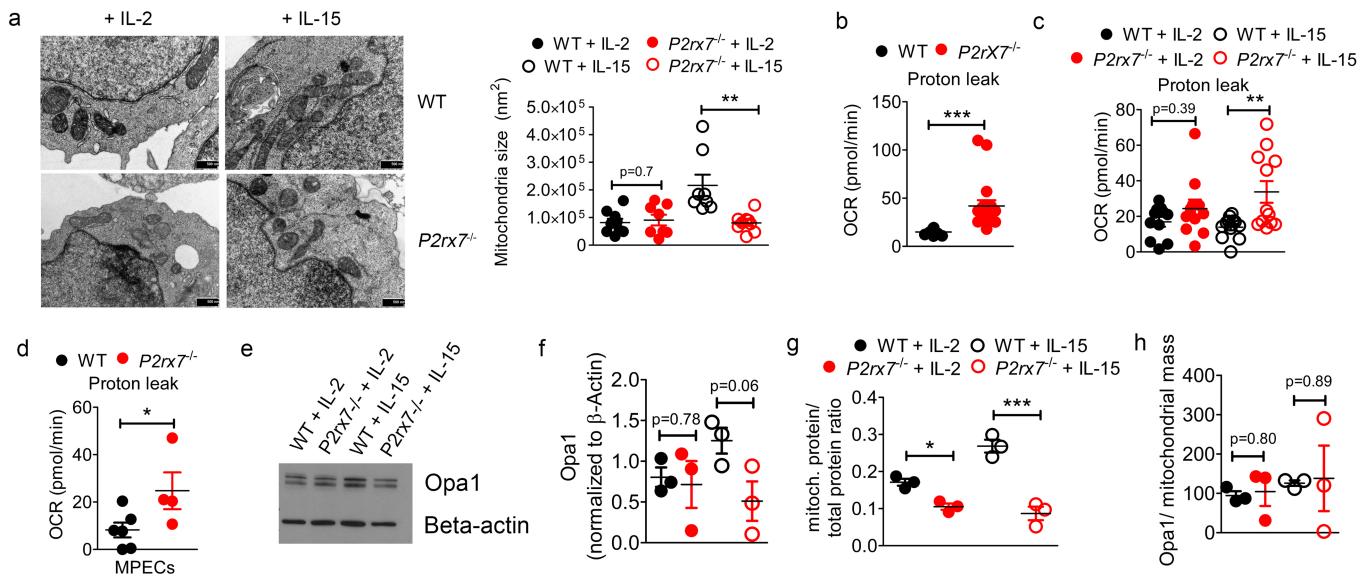
j, **k**, Congenically distinct wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells were stained with CFSE, mixed 1:1 and co-adoptively transferred into *Rag2*^{−/−} mice. Data from two independent experiments, $n = 6$ total. **j**, Ratio of *P2rx7*^{−/−} to wild-type P14 CD8⁺ T cells in the blood of *Rag2*^{−/−} hosts at indicated times post-transfer. **k**, Percentages of *P2rx7*^{−/−} and wild-type P14 CD8⁺ T cells proliferating over four cycles in spleens of *Rag2*^{−/−} hosts three weeks after transfer. All data shown as mean \pm s.e.m.

a–d, **f–h**, **k**, Two-tailed Student's *t*-test; **e**, two-way ANOVA with Bonferroni's post-test; $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$.



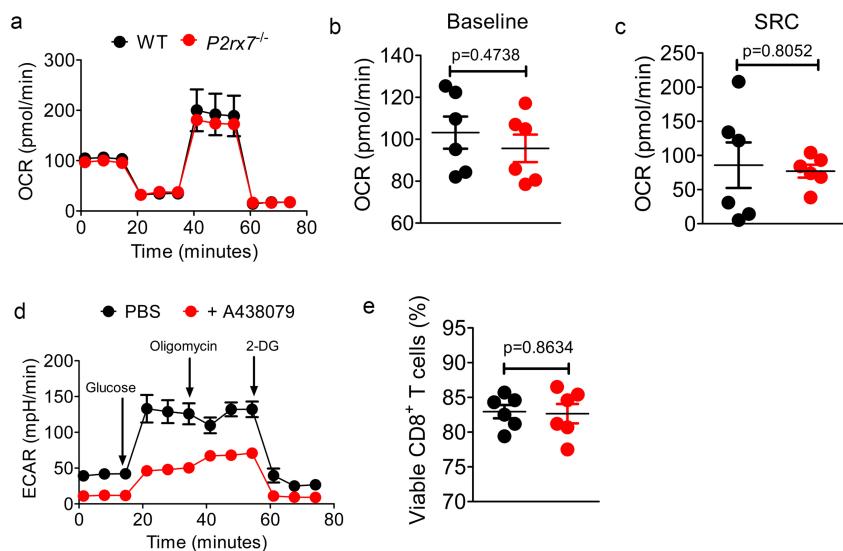
Extended Data Fig. 5 | Defective metabolism of IL-15-polarized CD8⁺ T cells in the absence of P2RX7. **a**, Viability of wild-type and P2rx7^{-/-} P14 cells maintained under culture conditions used for extracellular flux assays for 4 h (data from two independent experiments, $n = 11$ from all experiments). **b-d**, Wild-type or P2rx7^{-/-} P14 cells were activated in vitro and subsequently polarized in IL-2 or IL-15 for 72 h and assayed for ECAR to measure aerobic glycolysis (**b**); uptake of TMRE to measure mitochondrial membrane potential (data normalized to MTG staining; **c**); and staining with MTG to determine total mitochondrial mass (**d**). Data from three independent experiments, $n = 5-6$ from all experiments. **e**, Wild-type and P2rx7^{-/-} P14 cells were activated in vitro for 72 h,

then stimulated with IL-15 (or not) for 30 min and immediately assayed for pSTAT5 expression. Data are representative of two independent experiments ($n = 6$ total). **f**, Wild-type and P2rx7^{-/-} P14 cells were activated in vitro for 72 h, then stimulated with IL-15 or IL-2 for 72 h and assayed for expression of CD62L. Data are representative of three independent experiments ($n = 6$ total). **g**, Numbers of viable wild-type and P2rx7^{-/-} P14 cells following activation and subsequent culture in IL-15 or IL-2 for the indicated number of days. Values for wild-type and P2rx7^{-/-} cells were not significantly different ($P > 0.05$). Data from three independent experiments ($n = 3-6$ total). **a-d, g**, Mean \pm s.e.m.; **a, c, d, g**, two-tailed Student's *t*-test; ** $P \leq 0.01$.



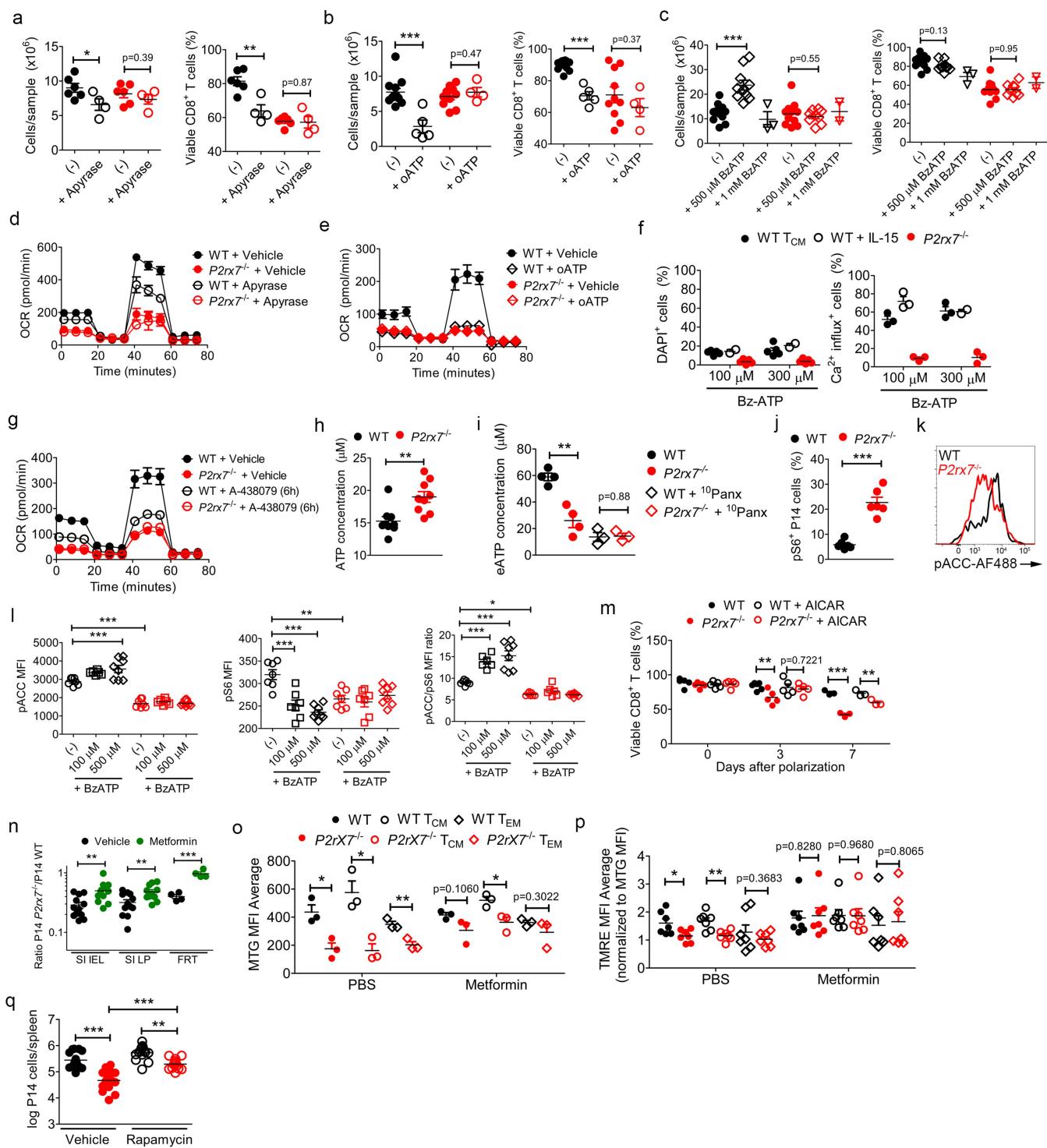
Extended Data Fig. 6 | P2RX7 controls mitochondrial integrity in CD8⁺ T cells during immune responses. **a**, Left, representative electron micrographs showing mitochondrial structures; right, mitochondrial area measurements of in vitro-activated wild-type and *P2rx7*^{-/-} P14 cells following culture for 72 h in IL-2 or IL-15. Representative of 3 independent experiments ($n = 8$ –9 in total). Black bars indicate 500 nm. **b, c**, Wild-type and *P2rx7*^{-/-} P14 cells were activated in vitro for 72 h and assayed at that time (**b**) or after a further 72 h of culture in IL-2 or IL-15 (as in Fig. 3b) (**c**) for OCR. Graphs show values for proton leak (the difference in OCR values after oligomycin and after antimycin A/rotenone addition; Fig. 3b). Data are from three independent experiments ($n = 11$ –18 total). **d**, Calculated proton leak derived from OCR measurements on in vivo activated wild-type and *P2rx7*^{-/-} P14 CD8⁺ MPECs described in Fig. 2d–f (data from three independent experiments, data pooled from five mice per experiment; $n = 4$ –6 wells in total). **e–h**, Wild-type

and *P2rx7*^{-/-} P14 cells were activated in vitro and polarized with either IL-2 or IL-15 (as in Fig. 3b); total cell (**e–g**) and mitochondrial (**g, h**) protein extracts were collected for protein quantification experiments. Data from three independent experiments, samples pooled from $n = 6$ mice total (2 mice per experiment). **e**, Representative blot showing Opa1 expression in polarized wild-type or *P2rx7*^{-/-} P14 cells, in comparison with β-actin (for gel source data, see Supplementary Fig. 1). **f**, Opa1 protein levels in polarized wild-type or *P2rx7*^{-/-} P14 cells, normalized to β-actin. **g**, Mitochondrial concentration (normalized by total protein concentration) in polarized wild-type or *P2rx7*^{-/-} P14 cells. **h**, Opa1 protein levels in polarized wild-type or *P2rx7*^{-/-} P14 cells, normalized to total mitochondrial concentration. **a–d, f–h**, Mean \pm s.e.m.; **a, b, d**, two-tailed Student's *t*-test; **c, f–h**, two-sided Mann–Whitney's test; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$.



Extended Data Fig. 7 | Normal metabolic function of naive $P2rx7^{-/-}$ P14 CD8 $^{+}$ T cells, while pharmacological inhibition of P2RX7 compromises aerobic glycolysis of in vitro-activated CD8 $^{+}$ T cells.
 a–c, Naive wild-type and $P2rx7^{-/-}$ P14 cells were isolated and evaluated for different metabolism parameters (data from three independent experiments, $n = 6$ total). d, e, Human CD8 $^{+}$ T cells were activated in

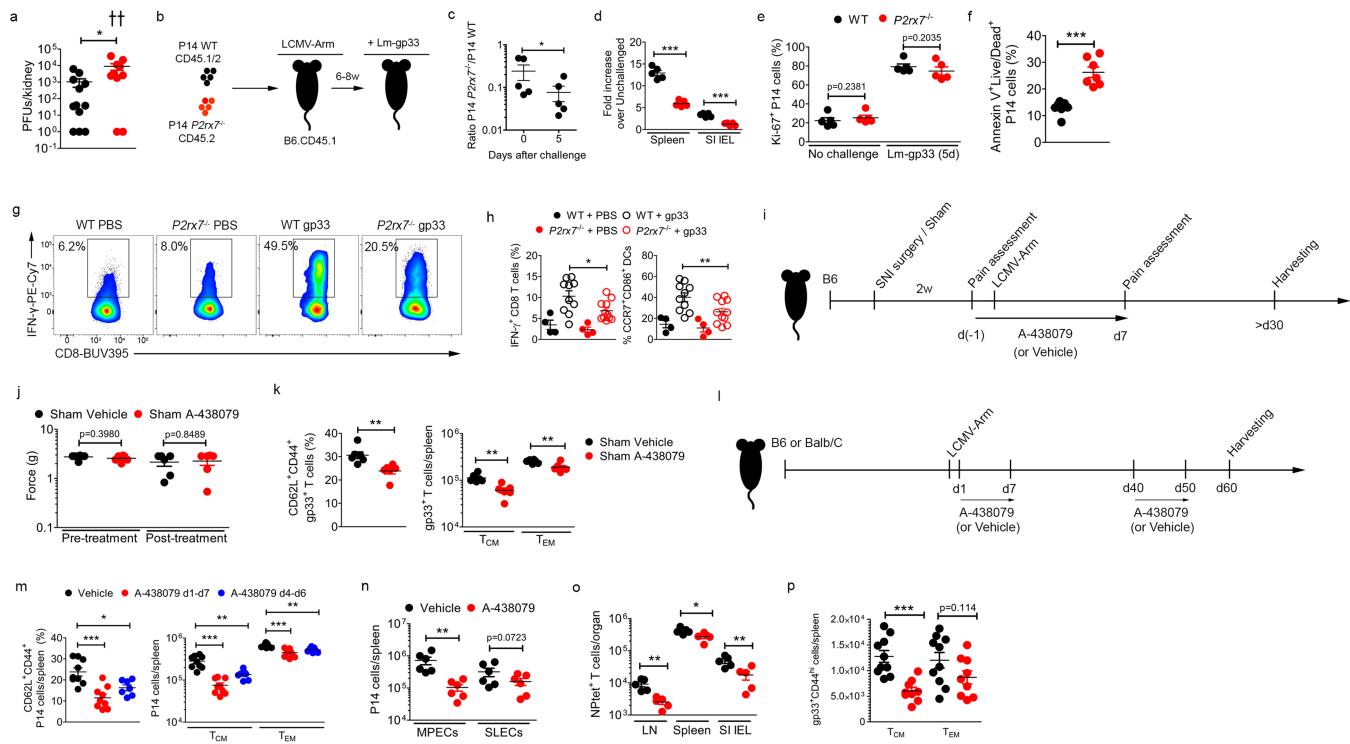
vitro, with the P2RX7 inhibitor A-438079 or vehicle control added 20 h after initiation of the culture, and assessed at 72 h for ECAR (**d**) or for viability of cells cultured in parallel under the same conditions as those used for extracellular flux assays (**e**). **d, e**, Data are from three independent experiments, $n = 4–6$ from all experiments. All data shown as mean \pm s.e.m.; **b, c, e**, Two-tailed Student's *t*-test.



Extended Data Fig. 8 | See next page for caption.

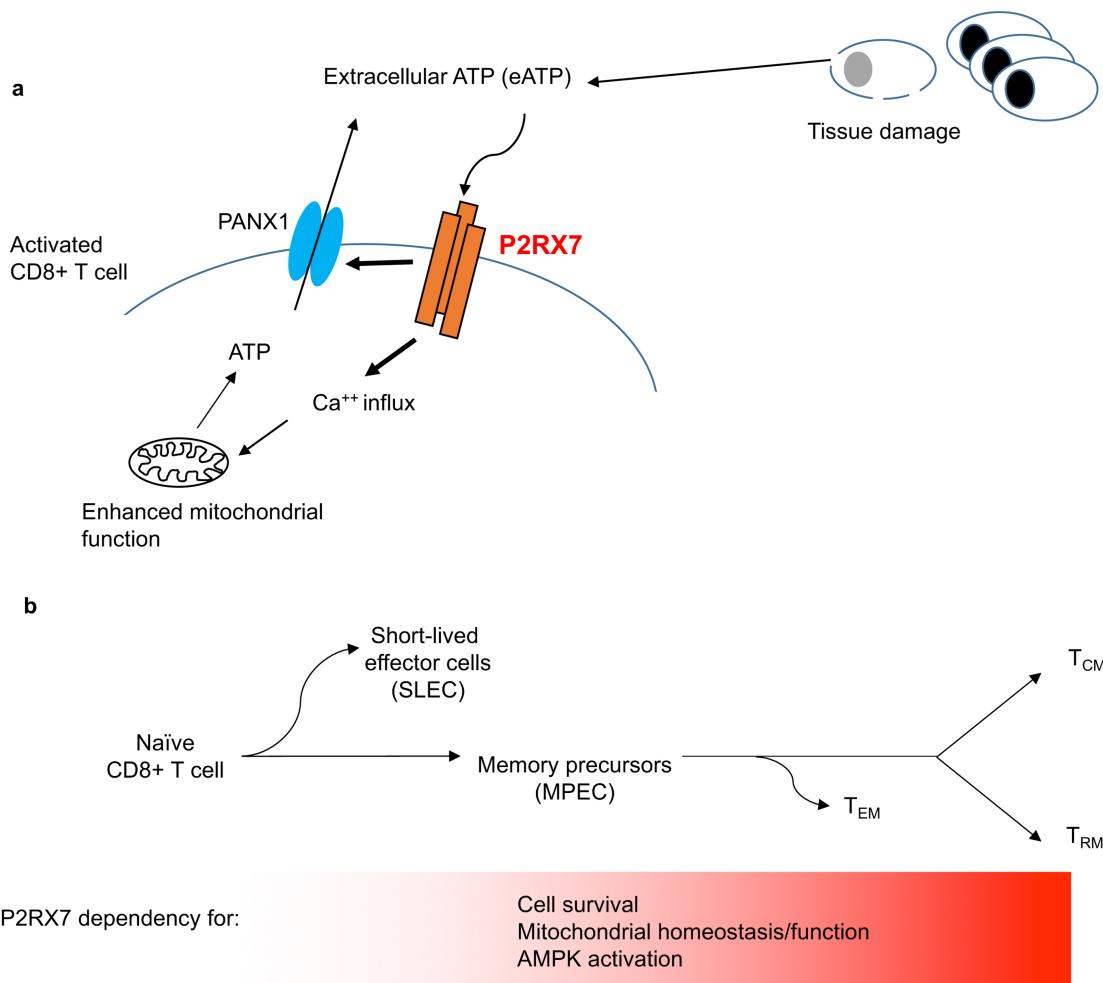
Extended Data Fig. 8 | P2RX7-mediated eATP sensing is crucial for optimal CD8⁺ T cell immunometabolism by regulation of the AMPK-mTOR pathway. **a–g.** Wild-type and *P2rx7*^{−/−} P14 cells were activated in vitro and polarized with IL-15 (as in Fig. 3b). Data from three independent experiments, samples pooled from $n = 6$ mice per experiment; $n = 3–12$ total samples. **a–c.** Numbers (left) and viability (right) of P14 cells in cultures supplemented with apyrase (**a**), oATP (**b**) or BzATP (**c**) during cell culture. **d, e, g.** IL-15-polarized wild-type or *P2rx7*^{−/−} P14 cells were assayed for OCR 1 h after addition of apyrase (**d**) or oATP (**e**), or 6 h after addition of A-438079 (**g**). **f.** IL-15-polarized cells or ex vivo wild-type P14 T_{CM} cells (isolated 4 weeks after LCMV infection) were incubated with DAPI (left) or Indo-1 (right) and stimulated with the indicated concentrations of BzATP during kinetic flow cytometric analysis. The percentage of cells showing DAPI uptake (left) or Ca²⁺ influx (right) over 30 min are shown. **f.** Data from two independent experiments, samples pooled from $n = 5$ mice total; $n = 2–5$ samples. **h.** In vitro-activated (72 h) wild-type and *P2rx7*^{−/−} P14 cells were assayed for intracellular ATP concentrations. Data from three independent experiments, $n = 9$ total. **i.** In vitro-activated, IL-15 polarized (24 h post-polarization) wild-type or *P2rx7*^{−/−} P14 cells were assayed for extracellular ATP concentration, following culture without or with the Panx1 inhibitor ¹⁰Panx. Data from two independent experiments, $n = 3–4$ total samples (pooled from six mice). **j.** Wild-type and *P2rx7*^{−/−} P14 cells were co-adoptively transferred and assayed 4 weeks after LCMV infection (as in Fig. 1a) and the ex vivo frequency of pS6-expressing cells was determined by flow cytometry. Data are from two independent experiments ($n = 6$ total). **k.** Expression of pACC in IL-15-polarized wild-type (black) and *P2rx7*^{−/−} (red) P14 cells (relative to Fig. 3i; representative of three independent experiments, $n = 6$ total). **l.** In vitro-activated and IL-15-polarized wild-type and *P2rx7*^{−/−}

P14 cells were cultured for 6 h with the indicated concentrations of BzATP, then stained for pACC (left) and pS6 (centre), and the pACC/pS6 ratio was determined (right). Data from three independent experiments, $n = 6–8$ total. **m.** In vitro-activated wild-type and *P2rx7*^{−/−} P14 cells were IL-15-polarized in the presence or absence of AICAR as in Fig. 3l. The percentage of viable cells at the indicated times following initiation of IL-15 culture with or without AICAR is indicated. Data are from three independent experiments ($n = 3–6$ total; samples pooled from $n = 6$ mice total). **n–p.** Wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells were mixed 1:1 and co-adoptively transferred into B6.SJL mice that were subsequently infected with LCMV, and donor cells identified as in Fig. 1. The mice were treated with metformin or PBS control during the first week of LCMV infection, and the cells were analysed at day 30. Data are compiled from three independent experiments ($n = 11–12$ total, $n = 4$ for FRT samples). **n.** Relates to Fig. 3m; ratio of *P2rx7*^{−/−} to wild-type P14 cells in the indicated non-lymphoid tissues ($n = 9$ except FRT, $n = 4$). **o, p.** Measurements of mitochondrial mass (measured using MTG) (**o**) and mitochondrial membrane potential (measured by TMRE staining, normalized to MTG staining) (**p**) for the indicated splenocyte subsets ($n = 3–6$ total samples). **q.** Wild-type and *P2rx7*^{−/−} P14 CD8⁺ T cells were mixed 1:1 and co-adoptively transferred into B6.SJL mice that were subsequently infected with LCMV, and donor cells were identified as in Fig. 1. The mice were treated with rapamycin or PBS control between days 4 and 8 post-LCMV infection, and the cells were analysed at day 30. The numbers of wild-type and *P2rx7*^{−/−} P14 cells are shown (log-transformed). Data are compiled from three independent experiments ($n = 15$ total). **a–j, l–q.** Mean \pm s.e.m.; **a–c, h–j, m–p,** two-tailed Student's *t*-test; **l, q,** one-way ANOVA with Tukey's post-test; $^*P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$.



Extended Data Fig. 9 | P2RX7 deficiency compromises memory CD8⁺ T cell function, and P2RX7 pharmacological blockade impairs generation of CD8⁺ T cell memory cells in vivo. **a**, Wild-type or *P2rx7*^{-/-} mice were infected with LCMV-Cl13, and kidney PFU levels were quantified 4 weeks later. Data from 3 independent experiments, $n = 7$ total. **b-f**, Wild-type and *P2rx7*^{-/-} P14 CD8⁺ T cells were mixed 1:1 and co-adoptively transferred into B6.SJL mice that were subsequently infected with LCMV and then challenged (or not) with Lm-gp33 6–8 weeks later (**b**). Data are from 2 independent experiments ($n = 5$ –7 total). **c**, Ratio of *P2rx7*^{-/-} to wild-type splenic P14 CD8⁺ T cells before (0 d) and after challenge (5 d). **d**, Fold increase in numbers of wild-type or *P2rx7*^{-/-} P14 cells in indicated tissues, relative to mice that did not receive Lm-gp33 challenge. **e**, Percentage of cells in cell cycle, determined by Ki-67 staining. **f**, Frequency of dying cells indicated by the percentage of annexin V⁺Live-Dead⁺ cells in mice 5 d after Lm-gp33 challenge. **g, h**, Wild-type and *P2rx7*^{-/-} P14 CD8⁺ T cells were individually transferred into B6.SJL mice that were subsequently infected with LCMV. After 6–8 weeks, the mice were transcervically challenged with gp33 or PBS as in Fig. 4d. **g**, Flow cytometry plots for IFN- γ production by wild-type or *P2rx7*^{-/-} P14 cells in mice treated with PBS or gp33. **h**, Percentage of IFN- γ bystander (non-P14) CD8⁺ T cells (left) and percentage of CCR7⁺CD86⁺ dendritic cells (right) in the FRT 12 h later. **g, h**, Data are from three independent experiments, $n = 4$ –11 total. **i**, Schematic of experimental scheme combining spared nerve injury (SNI), LCMV infection and A-438079 treatment. For surgery, two of the three branches of the sciatic nerve in one hind limb were exposed and cut (Fig. 4) or left uncut (sham). After 2 weeks mice were assayed for pain sensitivity, then

infected with LCMV with or without A-438079 treatment for the first week after infection. Mice were assayed again for pain sensitivity (day 7) and subsequent development of central (CD62L⁺) and effector (CD62L⁻) memory cells specific for the LCMV epitope gp33 (after day 30). Data are compiled from two independent experiments, $n = 6$ from all experiments. **j**, Pain sensitivity of sham-surgery mice (pre- and post-treatment). **k**, Percentages of gp33-specific T_{CM} cells (left) and numbers of gp33-specific T_{CM} and T_{EM} cells (right) in sham surgery animals. **l-p**, In other studies, B6 or Balb/C mice were adoptively transferred or not with wild-type P14 cells, infected with LCMV, and treated with A-438079 in the time frames indicated, relative to infection. Data from 2–3 independent experiments, $n = 5$ –10 total. **m**, Percentages of CD62L⁺CD44⁺ (T_{CM}) P14 cells per spleen (left) and spleen P14 T_{CM} and T_{EM} cell numbers (right) from the different treatment groups at 4 weeks post-infection. **n**, P14 recipient mice were treated with PBS or A-438079 for the first week following LCMV infection, then assayed at 3 weeks for numbers of MPECs and SLECs. **o**, Balb/C mice were infected with LCMV, treated with A-438079 throughout the first week post-infection, and the numbers of LCMV-specific (NPtet⁺) CD8⁺ T cells were quantified in the spleen, lymph nodes and SI-IEL at 4 weeks post-infection. **p**, B6 mice infected with LCMV were treated with A-438079 between days 40 and 50 post-infection and the numbers of gp33⁺ CD8⁺ T cells (T_{CM} and T_{EM}) per spleen were quantified at 8 weeks post-infection. Data shown as mean \pm s.e.m.; **a, c-f, k, n-p**, two-tailed Student's *t*-test; **j**, two-sided Mann-Whitney's test; **h, m**, one-way ANOVA with Tukey's post-test; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$.



Extended Data Fig. 10 | Model of P2RX7 function in regulating long-lived CD8⁺ T cell memory. **a**, After activation of CD8⁺ T cells, P2RX7 is stimulated by eATP (derived from damaged cells or exported from live activated cells). This induces calcium influx, increased mitochondrial metabolic activity and activation of the ATP export channel PANX1 (bold arrows). Generation of eATP through PANX1 sustains P2RX7 activation, further promoting mitochondrial function and T cell homeostasis. **b**, The effect of P2RX7 function is magnified as the CD8⁺ T cell response progresses into memory phase, with P2RX7 deficiency having little

effect on effector T cells, while severely compromising survival and metabolic function in memory T cells, impairing the generation of long-lived central memory (T_{CM}) and CD103^{high}CD69^{high} resident memory (T_{RM}) populations. P2RX7 activation is proposed to stimulate AMPK activation, which may arise from both calcium influx and increased AMP/ATP ratios as a result of P2RX7 and PANX1 activity. A requirement for eATP stimulation of P2RX7 persists throughout memory CD8⁺ T cell maintenance.

Targeting STING with covalent small-molecule inhibitors

Simone M. Haag^{1,4}, Muhammet F. Gulen^{1,4}, Luc Reymond², Antoine Gibelin², Laurence Abrami¹, Alexiane Decout¹, Michael Heymann¹, F. Gisou van der Goot¹, Gerardo Turcatti², Rayk Behrendt³ & Andrea Ablasser^{1*}

Aberrant activation of innate immune pathways is associated with a variety of diseases. Progress in understanding the molecular mechanisms of innate immune pathways has led to the promise of targeted therapeutic approaches, but the development of drugs that act specifically on molecules of interest remains challenging. Here we report the discovery and characterization of highly potent and selective small-molecule antagonists of the stimulator of interferon genes (STING) protein, which is a central signalling component of the intracellular DNA sensing pathway^{1,2}. Mechanistically, the identified compounds covalently target the predicted transmembrane cysteine residue 91 and thereby block the activation-induced palmitoylation of STING. Using these inhibitors, we show that the palmitoylation of STING is essential for its assembly into multimeric complexes at the Golgi apparatus and, in turn, for the recruitment of downstream signalling factors. The identified compounds and their derivatives reduce STING-mediated inflammatory cytokine production in both human and mouse cells. Furthermore, we show that these small-molecule antagonists attenuate pathological features of autoinflammatory disease in mice. In summary, our work uncovers a mechanism by which STING can be inhibited pharmacologically and demonstrates the potential of therapies that target STING for the treatment of autoinflammatory disease.

STING is an intracellular signalling molecule that senses cyclic dinucleotides from bacterial sources or cyclic GMP-AMP (2'-3')—produced by the cytosolic DNA sensor cyclic GMP-AMP synthase—and, upon stimulation, triggers the production of type I interferons (IFNs) and other inflammatory mediators^{1,3–10}. Chronic activation of STING has previously been implicated in the pathogenesis of monogenic autoinflammatory conditions, such as Aicardi–Goutières syndrome and STING-associated vasculopathy with onset in infancy^{11–14}. In addition, accumulating evidence suggests a pathogenic role for STING in a range of more complex inflammatory diseases^{15–20}. Thus, STING represents an attractive target for therapeutic intervention.

To discover molecules that inhibit STING, we performed a cell-based chemical screen and identified two nitrofuran derivatives—C-178 and C-176—that strongly reduced STING-mediated, but not RIG-I- or TBK1-mediated, IFN β reporter activity (Fig. 1a, b and Extended Data Fig. 1a–d). A limited structure–activity relationship analysis revealed that both the nitro group and the furan moiety were essential for the bioactivity of the compounds, and that substituents at the 4-position of the phenyl ring fine-tuned their inhibitory potency (Extended Data Fig. 1e, f). Studies in mouse bone marrow-derived macrophages (BMDMs) corroborated the finding that C-178 potently and selectively suppressed the STING responses elicited by distinct bona fide activators (Fig. 1c–e and Extended Data Fig. 2a, b). Profiling the effect of C-178 on the global transcriptional program initiated by the STING agonist 10-carboxymethyl-9-acridanone (CMA) in BMDMs showed a substantial reduction of 99.6% (498) of the 500 most-upregulated genes (Extended Data Fig. 2c, d). Of note, exposure to C-178 alone did not

appreciably affect the gene expression profile of BMDMs relative to DMSO-treated samples. In addition, C-178 inhibited the CMA-induced phosphorylation of TBK1, a key downstream protein kinase of STING (Fig. 1f). Notably, however, C-178 did not appreciably affect STING responses in human cells (Extended Data Fig. 2e). The species-specific activity of C-178 and C-176 suggested that the compounds directly target mouse STING (mmSTING) but not human STING (hsSTING).

5,6-Dimethylxanthenone-4-acetic acid (DMXAA) and CMA, which are mouse-specific activators of STING, have previously been described as acting through non-conserved amino acid residues located within the C-terminal region of mmSTING that encompass the ligand-binding domain^{21,22}. However, using chimaeric STING expression constructs we found that C-178 targeted the poorly characterized N-terminal portion of mmSTING that includes the transmembrane domains (Extended Data Fig. 2f). To map the amino acids involved in the inhibitory activity of C-178, we individually mutated several amino acids (alanine-scanning) located in the N-terminal part of mmSTING (amino acids 21–137) and screened for mutants with compromised sensitivity to C-178. A STING mutant with a Cys91-to-alanine replacement (STING(C91A)) was found to be entirely resistant to inhibition by C-178 (Fig. 2a, b). Based on the critical roles of both the nucleophilic cysteine residue on mmSTING and the electrophilic nitrofuran group of the compounds, we hypothesized that a covalent bond may be formed between C-178 and Cys91. Indeed, cellular washout experiments, native mass spectrometry assays on mmSTING (wild type and STING(C91S)) and top-down liquid-chromatography tandem mass spectrometry (LC-MS/MS) together confirmed that the Cys91 of mmSTING is covalently modified by C-178 and C-176 (Fig. 2c, Extended Data Fig. 3 and Supplementary Information).

To further characterize the binding of the compounds to mmSTING, we performed gel-based profiling studies using a derivative of C-176 with an installed alkyne group (C-176-AL), which is amenable to copper-catalysed azide alkyne cycloaddition (click chemistry) (Extended Data Fig. 4a, b). We found that within living cells C-176-AL effectively and specifically labelled mmSTING, whereas neither hsSTING nor mmSTING with a Cys91 substitution (either STING(C91A) or STING(C91S)) were targeted by the clickable compound (Fig. 2d, e and Extended Data Fig. 4c–e). We then used the gel-based protein profiling approach to study the degree of molecular selectivity in the covalent interaction between the compounds and STING. When compared to iodoacetamide azide, which is a non-specific cross-linking probe, an azide-based clickable C-176 probe (C-176-AZ) showed markedly lower background proteome reactivity (Extended Data Fig. 4f). Furthermore, testing STING against a panel of proteins that contain hyper-reactive cysteine residues revealed that only STING was efficiently labelled by C-176-AZ^{23,24} (Extended Data Fig. 4g). We also note that the installation of an additional methyl group at the central amine moiety of C-176 completely abolished the inhibitory capacity of the compound (Extended Data Fig. 1e, f). Collectively, these studies provide evidence that the nitrofuran scaffold does not randomly cross-react with cellular

¹Global Health Institute, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. ²Biomolecular Screening Facility, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. ³Institute for Immunology, Faculty of Medicine, Technical University Dresden, Dresden, Germany. ⁴These authors contributed equally: Simone M. Haag, Muhammet F. Gulen. *e-mail: andrea.ablasser@epfl.ch

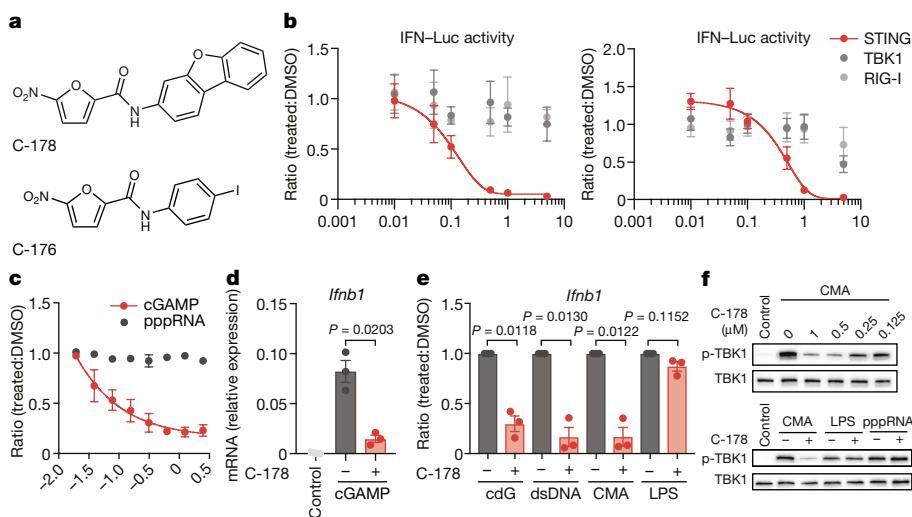


Fig. 1 | Identification of small-molecule inhibitors of STING. **a**, C-176 and C-178. **b**, IFN β luciferase reporter activity in HEK293T cells, transfected as indicated (compound concentration 0.01–5 μ M) ($n = 3$ biological replicates). **c**, Type I IFN bioassay of BMDMs with or without C-178 (0.02–2.5 μ M) ($n = 3$ biological replicates). cGAMP, cyclic GMP-AMP (2'-3'); pppRNA, 5'-triphosphate RNA. **d**, **e**, *Ifnb1* expression levels

proteins, but instead relies on a specific recognition event on STING that precedes the irreversible modification of Cys91. Moreover, the covalent bond between C-178 and STING may be formed by a nucleophilic addition of a nucleophilic side chain of Cys91 to the 4-position of the furan ring, followed by a subsequent intramolecular rearrangement (Fig. 2f and Extended Data Fig. 4h).

We next aimed to understand how the modification introduced by C-178 could antagonize STING. To activate TBK1, STING translocates from the endoplasmic reticulum to the Golgi apparatus^{1,25}. C-178, however, neither impaired the trafficking of STING to the

in BMDMs treated with C-178 (0.5 μ M) ($n = 3$ biological replicates). cdG, cyclic di-GMP; dsDNA, double-strand DNA; LPS, lipopolysaccharide. **f**, Immunoblot of phosphorylated TBK1 (p-TBK1) and TBK1 of BMDMs (one representative of $n = 3$ biological replicates). Data are mean \pm s.e.m. *P* values were calculated using two-tailed *t*-test. Nonlinear regression analysis is shown in **b**, **c**. For gel source data, see Supplementary Fig. 1.

Golgi apparatus nor affected the endolysosomal degradation of STING thereafter (Fig. 3a, b and Extended Data Fig. 5a, b). We then considered the possibility that C-178 may alter post-translational modifications of STING. We focused on palmitoylation, because this had previously been reported to rely on Cys91 and to be involved in the activation of TBK1²⁶. Indeed, we found that CMA-induced palmitoylation of STING was markedly attenuated in the presence of C-178, but that the palmitoylation of transferrin receptor or calnexin was not affected (Fig. 3c and Extended Data Fig. 5c–e). The exact mechanism through which palmitoylation regulates STING signalling is unclear,

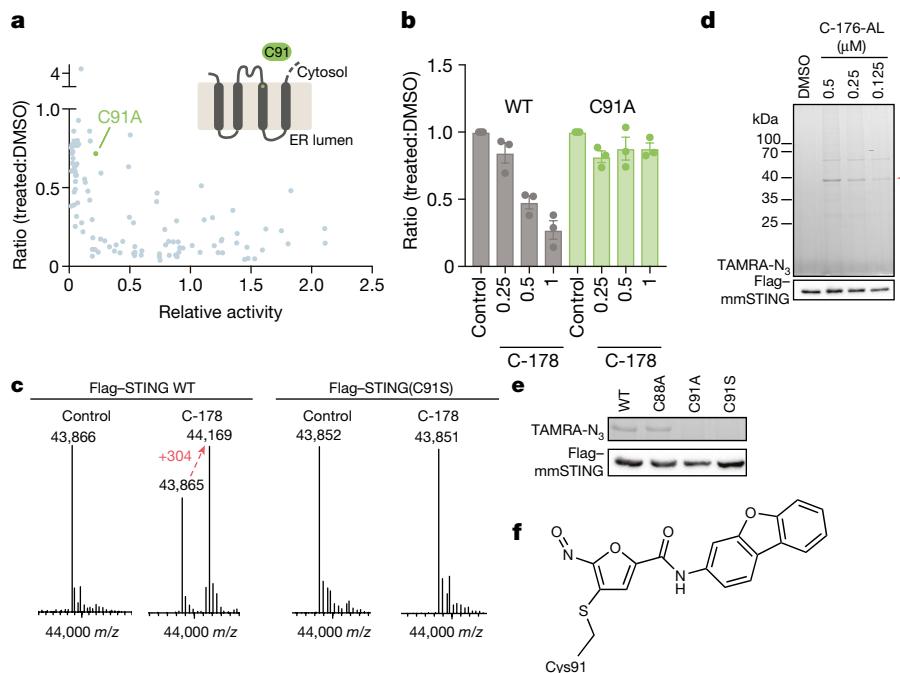


Fig. 2 | Inhibition of STING through covalent binding of C-178 to Cys91. **a**, Alanine scanning on mmSTING (amino acids 21–137). ER, endoplasmic reticulum. **b**, IFN β luciferase reporter measurements of HEK293T cells transfected with mmSTING constructs as indicated, with or without C-178. Data are mean \pm s.e.m. ($n = 3$ biological replicates). WT, wild type. **c**, Deconvoluted electrospray ionization mass spectra for Flag-mmSTING

purified from treated cells, as indicated. **d**, Gel showing labelling events of C-176-AL in HEK293T cells expressing Flag-mmSTING. TAMRA-N₃, tetramethylrhodamine azide. **e**, Gel showing labelling of Flag-mmSTING constructs by C-176-AL (0.25 μ M) in HEK293T cells. **f**, Proposed reaction adduct. One representative of $n = 2$ (**c**) or $n = 3$ (**d**, **e**) biological replicates. For gel source data, see Supplementary Fig. 1.

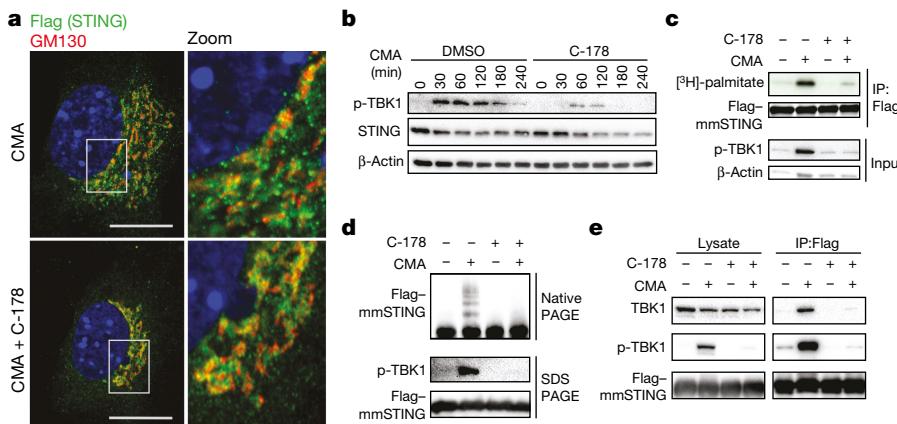


Fig. 3 | C-178 blocks palmitoylation-induced clustering of STING.

a, Staining of Flag (green), GM130 (red) and nuclei (DAPI) in mouse embryonic fibroblasts treated with CMA and expressing Flag-mmSTING (C-178 (1.5 μ M)). Scale bar, 20 μ m. **b**, Immunoblot of p-TBK1, STING and β -actin in mouse embryonic fibroblasts (C-178 (1 μ M)). **c**, [3 H]-palmitate incorporation of Flag-mmSTING in HEK293T cells and protein levels

(p-TBK1 and β -actin) after stimulation with CMA (2 h) (C-178 (1 μ M)). **d**, Blue native PAGE and SDS-PAGE of HEK293T cells with Flag-mmSTING after stimulation with CMA (C-178 (1 μ M)). **e**, Immunoprecipitation from Flag-mmSTING HEK293T cells and detection of TBK1 and p-TBK1 by immunoblotting. One representative of $n = 3$ biological replicates (**a–e**). For gel source data, see Supplementary Fig. 1.

as is the molecular event on STING that enables the recruitment of TBK1. We hypothesized that both of these processes may be connected and explain the mechanism through which C-178 inhibits STING. We observed that, after activation by CMA, STING assembles into clusters, which is a well-known effect of protein palmitoylation (Fig. 3d). Treatment of cells with C-178, C-176 or 2-bromopalmitate—which is a non-selective inhibitor of palmitoylation—completely prevented the formation of STING clusters (Fig. 3d and Extended Data Fig. 5f). Moreover, the clustering of STING paralleled the recruitment and phosphorylation of TBK1, which again was abrogated in the presence of C-178 (Fig. 3e and Extended Data Fig. 5g). Together, this suggests that palmitoylation triggers the assembly of a multimeric complex that enables STING to interact with TBK1. Moreover, C-178 interferes with this process by inhibiting the palmitoylation of STING (Supplementary Note 1).

We next studied the effects of pharmacological inhibition of STING in mice. Because we noticed improved solubility of C-176 relative to C-178, we chose this compound for in vivo studies. First we verified that the compounds target STING by using an in vivo click-chemistry approach and also assessed the pharmacokinetic profile of C-176 on single-dose intraperitoneal injection (Extended Data Fig. 6a, b). We next evaluated whether C-176 can suppress the induction of type I IFNs triggered by the administration of CMA. Of note, pretreatment with C-176 markedly reduced the CMA-mediated induction of serum levels of type I IFNs and IL-6 (Fig. 4a and Extended Data Fig. 6c). Thus, C-176 is effective in mice and—as expected for a covalent inhibitor—the short serum half-life does not limit its in vivo inhibitory capacity. To assess the potential of C-176 to antagonize STING in a model of autoinflammatory disease, we investigated its efficacy in *Trex1*^{−/−} mice. *Trex1*^{−/−} mice show signs of severe multi-organ inflammation caused by the persistent activation of the cyclic GMP-AMP synthase-STING pathway and recapitulate certain pathogenic features of Aicardi-Goutières syndrome in humans^{11,14,27–29}. Having verified that C-178 suppresses interferon-stimulated genes in cells from *Trex1*^{−/−} mice (Extended Data Fig. 7a), we performed a two-week in vivo efficacy study with C-176. Notably, treatment of *Trex1*^{−/−} mice with C-176 resulted in a significant reduction in serum levels of type I IFNs and in a strong suppression of inflammatory parameters in the heart (Extended Data Fig. 7b, c). Wild-type mice on a two-week treatment with C-176 showed no evident signs of overt toxicity (Extended Data Fig. 6d–g). We next conducted a three-month trial with C-176 in *Trex1*^{−/−} mice, which demonstrated marked amelioration of various signs of systemic inflammation (Fig. 4b, c and Extended Data Fig. 7e). Thus, C-176 attenuates STING-associated autoinflammatory disease in mice.

Given that the palmitate-modified cysteines are conserved residues and equally important for the functionality of hsSTING, we next searched for compounds that are capable of inhibiting the human protein. Further derivatization of the C-176 and C-178 scaffold yielded two structurally related compounds—C-170 and C-171—that efficiently inhibited both hsSTING and mmSTING through the same mechanism of action as that detailed above (Extended Data Fig. 8 and Supplementary Information). Encouraged by these results, we aimed to identify a more advanced covalent antagonist of hsSTING and performed another chemical screen. Progression of candidate hits through counter screens and validation screens identified H-151 (Fig. 5a). H-151 potently inhibited hsSTING, as evidenced by abrogation of type I IFN responses, reduction of TBK1 phosphorylation and suppression of hsSTING palmitoylation without affecting respective controls (Fig. 5b–e and Extended Data Fig. 9c). Using an alkyne analogue of

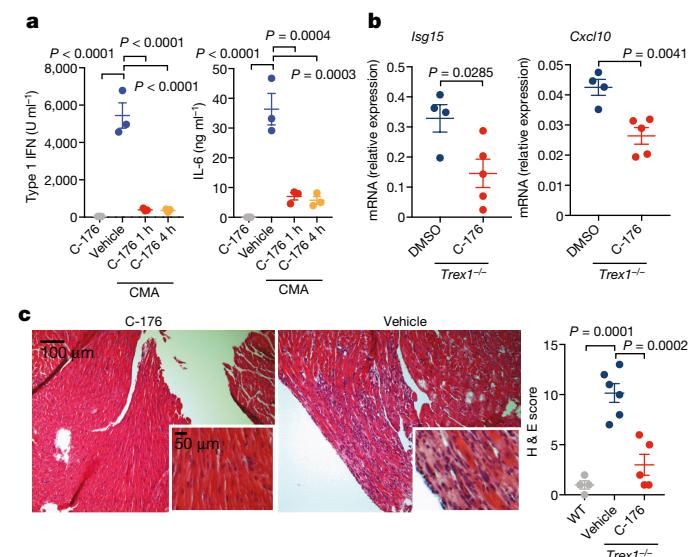


Fig. 4 | In vivo effects of C-176. **a**, Serum levels of type I IFNs and IL-6 from wild-type mice treated with C-176 (for 1 h and 4 h) and stimulated with CMA ($n = 3$ mice). **b**, **c**, Levels of *Isg15* and *Cxcl10* mRNA in the heart of *Trex1*^{−/−} mice (C-176 $n = 5$ and vehicle $n = 4$ mice) and heart histological analysis of wild-type ($n = 4$) or *Trex1*^{−/−} mice (C-176 $n = 5$ and vehicle $n = 6$ mice) treated for three months. H & E, haematoxylin and eosin. Data are mean \pm s.e.m. P values calculated using one-way ANOVA (**a**, **c**) or two-tailed *t*-test (**b**). Representative image shown in **c**. For source data, see Supplementary Table 1.

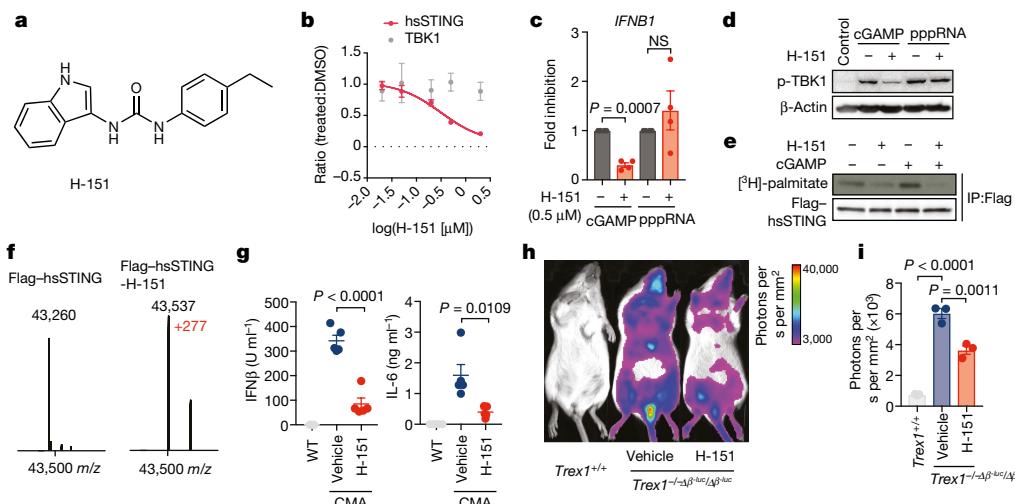


Fig. 5 | H-151 is a covalent STING antagonist. **a**, Structure of H-151. **b**, IFN β luciferase reporter measurements of HEK293T cells transfected as indicated (concentration of H-151 0.02–2 μ M). **c**, *IFNB1* mRNA levels in THP-1 cells pretreated with H-151 and stimulated as shown ($n = 4$). **d**, Immunoblots of p-TBK1 and β -actin of THP-1 cells stimulated for 2 h (H-151 0.5 μ M). **e**, [3 H]-palmitate incorporation of Flag–hsSTING after stimulation with cGAMP (3 h) (H-151, 1 μ M). **f**, Deconvoluted electrospray ionization mass spectra for Flag–hsSTING purified from treated cells as indicated. **g**, Serum levels of type I IFNs and IL-6 from

H-151 (H-151-AL), we observed a concentration-dependent interaction with hsSTING, with low background reactivity (Extended Data Fig. 9d–g). Mass spectrometry analysis of hsSTING purified from cells treated with H-151 revealed a mass shift that can be explained by an irreversible addition of H-151 to hsSTING (Fig. 5f). Mutagenesis studies together with top-down LC-MS/MS analysis confirmed that the irreversible modification introduced by H-151 to hsSTING again depended on Cys91 (Extended Data Fig. 9h–j and Supplementary Information). We next investigated the bioactivity of the H-151 compound in mice (Extended Data Fig. 10a, b). After intraperitoneal administration, H-151 reached effective systemic levels, displayed a short half-life in the serum and formed an adduct to mmSTING (Extended Data Fig. 10c, d). Notably, pretreatment of H-151 markedly reduced systemic cytokine responses in CMA-treated mice (Fig. 5g). Moreover, H-151 exhibited notable efficacy in *Trex1*^{−/−} mice that expressed a bioluminescent IFN β reporter, when administered for one week (Fig. 5h, i). Taken together, these data show that H-151 is a highly potent and selective small-molecule antagonist of STING that has noteworthy inhibitory activity both in human cells and in vivo.

In summary, we report the discovery of small-molecule inhibitors of STING that exploit an unanticipated covalent mechanism of action (Extended Data Fig. 11). Moreover, our study provides proof-of-concept that STING antagonists are efficacious in the treatment of autoinflammatory disease. The possibility of being able to pharmacologically antagonize STING may advance our understanding of its relevance in various contexts of health and disease.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0287-8>.

Received: 8 September 2017; Accepted: 24 May 2018;

Published online 4 July 2018.

- Ishikawa, H., Ma, Z. & Barber, G. N. STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature* **461**, 788–792 (2009).
- Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP–AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science* **339**, 786–791 (2013).
- Burdette, D. L. et al. STING is a direct innate immune sensor of cyclic di-GMP. *Nature* **478**, 515–518 (2011).
- Ablässer, A. et al. cGAS produces a 2'-5'-linked cyclic dinucleotide second messenger that activates STING. *Nature* **498**, 380–384 (2013).
- Diner, E. J. et al. The innate immune DNA sensor cGAS produces a noncanonical cyclic dinucleotide that activates human STING. *Cell Reports* **3**, 1355–1361 (2013).
- Gao, P. et al. Cyclic [G(2',5')pA(3',5')p] is the metazoan second messenger produced by DNA-activated cyclic GMP–AMP synthase. *Cell* **153**, 1094–1107 (2013).
- Wu, J. et al. Cyclic GMP–AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA. *Science* **339**, 826–830 (2013).
- Zhang, X. et al. Cyclic GMP–AMP containing mixed phosphodiester linkages is an endogenous high-affinity ligand for STING. *Mol. Cell* **51**, 226–235 (2013).
- Abe, T. & Barber, G. N. Cytosolic-DNA-mediated, STING-dependent proinflammatory gene induction necessitates canonical NF- κ B activation through TBK1. *J. Virol.* **88**, 5328–5341 (2014).
- Barber, G. N. STING: infection, inflammation and cancer. *Nat. Rev. Immunol.* **15**, 760–770 (2015).
- Gall, A. et al. Autoimmunity initiates in nonhematopoietic cells and progresses via lymphocytes in an interferon-dependent autoimmune disease. *Immunity* **36**, 120–131 (2012).
- Jeremiah, N. et al. Inherited STING-activating mutation underlies a familial inflammatory syndrome with lupus-like manifestations. *J. Clin. Invest.* **124**, 5516–5520 (2014).
- Liu, Y. et al. Activated STING in a vascular and pulmonary syndrome. *N. Engl. J. Med.* **371**, 507–518 (2014).
- Crow, Y. J. & Manel, N. Aicardi–Goutières syndrome and the type I interferonopathies. *Nat. Rev. Immunol.* **15**, 429–440 (2015).
- Ahn, J., Gutman, D., Sajio, S. & Barber, G. N. STING manifests self DNA-dependent inflammatory disease. *Proc. Natl. Acad. Sci. USA* **109**, 19386–19391 (2012).
- Ahn, J. et al. Inflammation-driven carcinogenesis is mediated through STING. *Nat. Commun.* **5**, 5166 (2014).
- An, J. et al. Expression of cyclic GMP–AMP synthase in patients with systemic lupus erythematosus. *Arthritis Rheumatol.* **69**, 800–807 (2017).
- King, K. R. et al. IRF3 and type I interferons fuel a fatal response to myocardial infarction. *Nat. Med.* **23**, 1481–1487 (2017).
- Zeng, L. et al. ALK is a therapeutic target for lethal sepsis. *Sci. Transl. Med.* **9**, eaan5689 (2017).
- Kerur, N. et al. cGAS drives noncanonical-inflammasome activation in age-related macular degeneration. *Nat. Med.* **24**, 50–61 (2018).
- Cavilar, T., Deimling, T., Ablässer, A., Hopfner, K. P. & Hornung, V. Species-specific detection of the antiviral small-molecule compound CMA by STING. *EMBO J.* **32**, 1440–1450 (2013).
- Conlon, J. et al. Mouse, but not human STING, binds and signals in response to the vascular disrupting agent 5,6-dimethylxanthenone-4-acetic acid. *J. Immunol.* **190**, 5216–5225 (2013).
- Dennehy, M. K., Richards, K. A., Wernke, G. R., Shyr, Y. & Liebler, D. C. Cytosolic and nuclear protein targets of thiol-reactive electrophiles. *Chem. Res. Toxicol.* **19**, 20–29 (2006).

24. Weerapana, E. et al. Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature* **468**, 790–795 (2010).
25. Saitoh, T. et al. Atg9a controls dsDNA-driven dynamic translocation of STING and the innate immune response. *Proc. Natl Acad. Sci. USA* **106**, 20842–20846 (2009).
26. Mukai, K. et al. Activation of STING requires palmitoylation at the Golgi. *Nat. Commun.* **7**, 11932 (2016).
27. Stetson, D. B., Ko, J. S., Heidmann, T. & Medzhitov, R. Trex1 prevents cell-intrinsic initiation of autoimmunity. *Cell* **134**, 587–598 (2008).
28. Gao, D. et al. Activation of cyclic GMP–AMP synthase by self-DNA causes autoimmune diseases. *Proc. Natl Acad. Sci. USA* **112**, E5699–E5705 (2015).
29. Gray, E. E., Treuting, P. M., Woodward, J. J. & Stetson, D. B. Cutting edge: cGAS is required for lethal autoimmune disease in the Trex1-deficient mouse model of Aicardi–Goutières Syndrome. *J. Immunol.* **195**, 1939–1943 (2015).

Acknowledgements We thank N. Jordan, E. Simeoni and L. Muhandes for technical assistance and advice. We acknowledge the staff of the BSF-ACCESS screening platform at the EPFL for support, especially M. Champon and J. Bortoli, and the staff of the ISIC Mass Spectrometry platform at the EPFL, especially N. Gasilova. We thank the following core facilities for their support: BIOP, CPG and HCF. A.A. received grants from the SNF (BSSGI0-155984, 31003A_159836), the Geber Rüf Foundation (GRS-059_14) and the Else Kröner Fresenius Stiftung (2014_A250). R.B. received funding from an AGS Research Award and the German Research council (DFG - BE 5877/2-1).

Reviewer information *Nature* thanks Z. Chen, T. Taguchi, H. Wu and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.M.H., M.F.G., L.R., L.A., A.D. and M.H. designed, performed and analysed experiments, and S.M.H. and M.F.G. assisted in writing the manuscript. L.R. and A.G. synthesized chemical compounds. R.B. and M.F.G. designed, performed and analysed animal experiments. G.T. and G.F.v.d.G. provided advice. A.A. designed, performed and analysed experiments, wrote the manuscript with input from all authors, conceived the idea and supervised the study.

Competing interests A.A. is a consultant to IFM Therapeutics, LLC. A.A., S.M.H., L.R. and the EPFL have filed provisional patent applications related to STING inhibitors. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0287-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0287-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cells and cell culture conditions. Cells were cultured under 5% CO₂ and ambient O₂ at 37 °C in DMEM containing 10% fetal bovine serum (FBS), 100 IU ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin. BMDMs were generated by culturing bone marrow cells from wild-type mice in L929-conditioned medium. Mouse embryonic fibroblasts were generated according to standard conditions. LL171 cells and their use have previously been described³⁰. WI-38 cells and were purchased from ATCC. THP-1 cells were provided by A. Rösen-Wolff (University Hospital Dresden) and cultured in RPMI-1640 containing 10% FBS, 100 IU ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin. HEK293T cells were from D. Trono (EPFL). Cell lines were repeatedly tested for mycoplasma by PCR. No method of cell line authentication was used.

Mice and in vivo studies. C57BL/6J mice (stock number 000664) were purchased from Jackson Laboratories. TREX1-deficient mice were a gift from T. Lindahl³¹ and were backcrossed for >10 generations to C57BL/6N. Mice were maintained under specific-pathogen-free (SPF) conditions at EPFL. For the pharmacokinetic studies, wild-type mice were injected intraperitoneally with 750 nmol C-176 per mouse in 200 µl corn oil (Sigma). Blood was collected at 30 min, 2 h and 4 h and serum C-176 levels were measured by mass spectrometry (liquid chromatography-high-resolution mass spectrometry). To assess the in vivo inhibitory effect of C-176, wild-type mice (8–12 weeks of age) were injected either with vehicle or C-176. After 1 h or 4 h, CMA was administered at a concentration of 224 mg kg⁻¹. Four hours later, mice were euthanized and the serum was collected to measure CMA-induced cytokine levels. To assess the in vivo inhibitory effect of H-151, wild-type mice were injected intraperitoneally with 750 nmol H-151 per mouse in 200 µl 10% Tween-80 in PBS. After 1 h CMA (112 mg kg⁻¹) was administered, and after 4 h mice were euthanized and the serum was collected. The efficacy study in *Trex1*^{-/-} mice was conducted as follows: mice (2–5 weeks of age) were injected with 7.5 µl of C-176 or DMSO dissolved in 85 µl corn oil twice per day for 11 consecutive days. Mice were euthanized by anaesthetization in a CO₂ chamber followed by cervical dislocation. For toxicology studies, 8-week-old mice were injected daily with 562.5 nmol of C-176 for 2 weeks. At day 14, blood samples were collected in lithium-heparin-coated tubes (Microvette CB 300 Hep-Lithium), and plasma was isolated after centrifugation at 4 °C and then stored at –80 °C. Plasma parameters were measured using DimensionXpand Plus (Siemens Healthcare Diagnostics AG). For the peripheral blood cell profile, 100 µl of blood was collected in EDTA-K-coated tubes (Microvette CB 300 EDTA K2). Complete blood counts were analysed with an ADVIA120 haematology system (Siemens Healthcare Diagnostics AG). For the detection of luciferase activity, *Trex1*^{-/-} *Ifnb1*^{Δβ-luc/Δβ-luc} reporter mice³² (aged 4–7 weeks) were injected intraperitoneally daily for 7 days with 750 nmol H-151 or DMSO in 200 µl PBS 0.1% Tween-80. For in vivo imaging, mice were anaesthetized with isofluran and injected intravenously with 15 mg kg⁻¹ XenoLight p-luciferin (Perkin Elmer) in isotonic sodium chloride. Photon flux was quantified two minutes after injection on an In-vivo Xtreme II imaging device (Bruker) with binning set to 8 × 8 pixels and an integration time of 3 min. Animal experiments were approved either by the Service de la Consommation et des Affaires Vétérinaires of the canton of Vaud (Switzerland) or by the Landesdirektion Dresden (Germany) and were performed in accordance with the respective legal regulations.

Cell-based IFN-β promoter-reporter luciferase measurements. HEK293T cells were seeded in a 96-well plate and transfected using GeneJuice (Millipore) with a IFN-β promoter-reporter plasmid (pIFN-β-Gluc) in combination with indicated expression constructs. After 16 h, gaussia luciferase activity was measured in the supernatants using coelenterazine (PJK GmbH) as substrate.

High-throughput chemical compound screen. HEK293T cells that expressed mouse STING with an N-terminal mCherry tag³⁰ were transfected using GeneJuice (Millipore) with a construct encoding for cyclic di-GMP synthase in conjunction with an IFN-β firefly luciferase reporter plasmid. Three hours later, transfected cells were seeded in 384-well plates (Corning) coated with library compounds. For each compound, a 40-nl volume was selected to obtain a concentration of 10 µM in the assay plates. The amount of DMSO in each well was normalized to 0.1%. After overnight treatment, cells were lysed in lysis buffer (25 mM Tris-phosphate (pH 7.8), 2 mM DTT, 2 mM 1,2-diaminocyclohexane-N,N,N',N'-tetraacetic acid, 10% glycerol, 1% Triton X-100) for 20 min followed by the addition of firefly luciferase substrate. Reporter activity was measured using a Tecan Infinite plate reader. The screen was performed on ~20,000 compounds from a chemically diverse compound collection available at the BSF core facility at EPFL. For the identification of hsSTING-specific compounds, HEK293T cells expressing a human STING construct were transfected with a construct encoding mouse cyclic GMP-AMP synthase in conjunction with an IFN-β firefly luciferase reporter plasmid. Further analysis was performed as described above. The screen was performed on ~30,000 compounds from a chemically diverse compound collection available at the BSF core facility at EPFL.

Histological analyses. Tissues were fixed in 4% formaldehyde solution (SAV LP), embedded in paraffin, and sections were stained with haematoxylin (Merck) and eosin (Shandon) (H & E). Five individual sections of the same heart were scored for levels of tissue damage (0–3) and inflammation (0–3). Scores represent the sum per organ.

Type I IFN bioassay and enzyme-linked immunosorbent assay. Levels of mouse type I IFN were determined by incubating LL171 cells that stably express an ISRE-luciferase construct with supernatants or serum for 5 h. LL171 cells were lysed in passive lysis buffer (Promega) and luciferase activity was measured using luciferin as substrate. Levels of mouse IL-6 and human IP-10 were determined by enzyme-linked immunosorbent assay (BD Bioscience) according to the manufacturer's instructions.

Stimulation of cells. BMDMs (1 × 10⁶ cells ml⁻¹) were pretreated with DMSO, C-178 (0.5 µM unless otherwise indicated) or H-151 (0.5 µM) for 1 h, followed by stimulation with either CMA (250 µg ml⁻¹, Sigma), dsDNA (90-mer, 1.33 µg ml⁻¹), cyclic di-GMP or cGAMP (1.5 µg ml⁻¹, Biolog). Triphosphate RNA (166 ng ml⁻¹) or LPS (1 µg ml⁻¹, Invivogen) were used as controls. THP-1 cells (1 × 10⁶ cells ml⁻¹) were differentiated with PMA (100 ng ml⁻¹, Sigma) for at least 3 h and treated with C-170 (0.5 µM) or H-151 (0.5 µM or as indicated) for 2 h, followed by stimulation with cyclic GAMP (375 ng ml⁻¹) or triphosphate RNA (133 ng ml⁻¹) for 2–3 h (for mRNA expression analysis and p-TBK1 immunoblot) or overnight (for IP-10 production). Alternatively, WI-38 cells (0.15 × 10⁶ cells ml⁻¹) and THP-1 (1 × 10⁶ cells ml⁻¹) cells were pretreated with C-178 (0.5 µM) and stimulated with cGAMP (1.5 µg ml⁻¹) for 2–3 h. To transfect dsDNA, triphosphate RNA and cyclic dinucleotides, Lipofectamine 2000 (Life Technologies) was used. The sequence of the sense strand of the 90-mer DNA is as follows: 5'-TACAGAT CTACTAGTGTATCTATGACTGATCTGTACATGATCTACATACAGATCTACT AGTGTATCTATGACTGATCTGTACATGATCTACTA-3'.

Mutagenesis PCR and alanine scanning. Point mutants were generated by site-directed PCR mutagenesis via the Quikchange Primer Design method (Agilent) using PrimeSTAR Max DNA Polymerase (Takara) and suitable primers. Alanine scanning was performed as follows: HEK293T cells were plated in 96-well plates, transfected with plasmids encoding individual mmSTING alanine point-mutants together with an IFN-β luciferase reporter using GeneJuice (Merck Millipore), and then treated with C-178. Luciferase activity was assessed after overnight incubation and normalized to values obtained for wild-type mmSTING.

RNA sequencing. BMDMs were pretreated with C-178 for 1 h, followed by 2 h of stimulation with either DMSO or CMA. After CMA stimulation, RNA was isolated using RNeasy Mini kit (Qiagen). mRNA sequencing libraries were prepared using the TruSeq mRNA stranded LT (Illumina kit). Samples were sequenced by the NextSeq 500 system sequencing with 1 × 75 cycle ('single read', 'high output' mode). Resultant data files were converted to fastq format, demultiplexed into constituent libraries and trimmed. Short reads were aligned to the mouse genome mm10. Library preparation, RNA sequencing and differential gene expression analysis was performed by Microsynth (Switzerland). Heat maps of normalized values were generated using the web-based platform Automated Single-cell Analysis Pipeline³³.

Western blot analysis. Cells were lysed in 2× Laemmli buffer, immobilized protein on beads in sample reducing buffer followed by denaturing at 95 °C for 5 min. Cell lysates were separated by SDS-PAGE and transferred onto PVDF membranes. Blots were incubated with anti-STING (D2P2F), phospho-TBK1 (D52C2), TBK1 (D1B4) (all Cell Signaling Technology), anti-transferrin receptor (Thermo Scientific, 136800) anti-calsevelin (Millipore, MAB3126) or anti-Flag M2 (Sigma, F3165). As secondary antibodies, anti-rabbit-IgG-HRP or anti-mouse-IgG-HRP (1:2000) (Santa Cruz Biotechnology) were used. Anti-β-actin (C4, Santa Cruz, 1:5000) was used as control. ECL signal was recorded on the ChemiDoc XRS Biorad Imager and data were analysed with Image Laboratory (Biorad).

Quantitative RT-qPCR. Total RNA was isolated using the RNAeasy Mini Kit (Qiagen) and cDNA was synthesized using the RevertAid First Strand cDNA Synthesis kit (Fermentas). Quantitative RT-qPCR was performed in duplicates using Maxima SYBR Green Master Mix (Thermo Scientific) on a QuantStudio 5 machine. GAPDH was used as an endogenous normalization control to obtain relative expression data. Primer sequences are as follows: mmGapdh forward, 5'-GTCATCCCAGAGCTGAACG-3'; mmGapdh reverse, 5'-TCATACTTGGCAGGTTCTCC-3'; mmIfnb1 forward, 5'-CTCCA GCTCCAAGAAAGGAC-3'; mmIfnb1 reverse, 5'-TGGCAAAGGCAGTGTACAC TC-3'; mmTnf forward, 5'-TATGGCCCAGACCTTCACA-3'; mmTnf reverse, 5'-GGAGTAGACAAGGTACAAACCCATC-3'; mmIsg15 forward, 5'-AAGAAGC AGATTGCCAGAA-3'; mmIsg15 reverse, 5'-TCTGCGTCAGAAAGACCTCA-3'; mmCxcl10 forward, 5'-AAGTGCCTGGCTCATTTCTC-3'; mmCxcl10 reverse, 5'-GTGGCAATGATCTAACACG-3'; hsGAPDH forward, 5'-GAGTCACG GATTGGTCGT-3'; hsGAPDH reverse, 5'-GACAAGCTCCGTTCTCAG-3'; hsIFNB1 forward, 5'-CAGCATCTGCTGGTTGAAGA-3'; hsIFNB1 reverse,

5'-CATTACCTGAAGGCCAAGGA-3'; hsTNF forward, 5'-CCCGAGT GACAAGCCTGTAG-3'; hsTNF reverse, 5'-TGAGGTACAGGCCCTTGAT-3'. **Metabolic labelling with [³H]-palmitate.** Indicated Flag-STING constructs were expressed in HEK293T cells. For metabolic labelling, cells were starved for 1 h in Glasgow minimal essential medium buffered with 10 mM Hepes, pH 7.4 with C-178 or C-176 (1 μ M). Cells were then incubated for 2 h in IM with 200 μ Ci ml⁻¹ [³H]-palmitic acid (9,10-³H(N)) (American Radiolabelled Chemicals) in presence of C-178 or C-176, and with or without stimulation with CMA (250 μ g ml⁻¹). For immunoprecipitation, cells were washed three times in PBS, lysed for 30 min at 4°C in the following buffer (0.5% Nonidet P-40, 500 mM Tris pH 7.4, 20 mM EDTA, 10 mM NaF, 2 mM benzamidin and protease inhibitor cocktail (Roche)) and centrifuged for 3 min at 5000 r.p.m. Supernatants were incubated overnight at 4°C with the appropriate antibodies (anti-Flag, anti-transferrin receptor and anti-calnexin) and G sepharose beads (GE Healthcare, 17-0618-01). For radiolabelling experiments, after immunoprecipitation washed beads were incubated for 5 min at 90°C in reducing sample buffer before 4–20% gradient SDS-PAGE. Following SDS-PAGE, gels were incubated in a fixative solution (25% isopropanol, 65% H₂O, 10% acetic acid) and incubated for 30 min with signal enhancer Amplify NAMP100 (GE Healthcare). The radiolabelled products were revealed using autoradiography and quantified using the Typhoon Imager (ImageQuanTool, GE Healthcare).

Competition assay. HEK293T cells expressing Flag-STING were incubated with the indicated compounds and after 1 h, C-176-AL was added for 1 h. Cells were collected in PBS and analysed by in-gel analysis of C-176-AL-mediated labelling of STING (see 'Gel-based analysis of compound binding to STING').

Immunoprecipitation. Flag-STING expression was induced in HEK293T cells overnight by doxycycline (Sigma). Cells were incubated with or without C-178 or C-176 (1 μ M) for 1 h and treated with DMSO or CMA (250 μ g ml⁻¹) for 2 h. Cells were washed in PBS and lysed in lysis buffer (50 mM HEPES, 150 mM NaCl, 10% glycerin, 1 mM MgCl, 1 mM CaCl, 1% Brij-58 and protease inhibitor cocktail (Sigma P8340)) for 30 min. Flag-STING was immunoprecipitated using anti-Flag M2 affinity gel agarose gel (Sigma) for 2 h at 4°C. After stringent washing in lysis buffer and PBS, the supernatant was completely removed and the resin was boiled in sample buffer before SDS-PAGE was performed. For immunoprecipitation of endogenous STING, splenocytes were lysed in the above-mentioned lysis buffer and incubated with anti-STING (RD System AF6516) and G sepharose beads (GE Healthcare, 17-0618-01) overnight. Beads were washed in PBS and gel-based analysis of C-176-AL binding to STING was performed.

Gel-based analysis of compound binding to STING. HEK293T cells expressing Flag-STING were incubated with C-176-AL, C-175-AZ, iodoacetamide azide (Thermo Fisher) or H-151-AL in serum-free medium, collected in PBS and lysed by repetitive freezing and thawing. Forty-three microlitres of lysed cells was treated with a freshly prepared 'click reagent' mixture containing tris(benzyl-triazolylmethyl)amine (TBTA) (3 μ l per sample, 3 mM in 1:4 DMSO:*t*-ButOH), tetramethylrhodamine (TAMRA) azide (Thermo Fisher), SiR azide (Spirochrome) or SiR alkyne (Spirochrome) (2 μ l per sample, 1.25 mM in DMSO), and freshly prepared CuSO₄ (1 μ l per sample) and tris-(2-carboxyethyl)phosphine hydrochloride (TCEP) (1 μ l per sample) and incubated at room temperature for 30 min. The reaction was quenched by addition of reducing sample buffer. In-gel fluorescence was visualized using Fusion FX (Vilber Lourmat) and analysed by Fusion capt advance acquisition software.

Crosslinking with disuccinimidyl suberate. HEK293T cells expressing Flag-mmSTING were incubated with or without C-176 (1 μ M) for 1 h and treated with DMSO or CMA (250 μ g ml⁻¹) for 2 h. Crosslinking was performed in PBS with 1 mM disuccinimidyl suberate (DSS) (Thermo Fisher) freshly prepared in DMSO at room temperature for 1 h.

Blue native gel assay. Flag-STING expression was induced in HEK293T cells overnight by doxycycline. Cells were washed, collected and lysed using the native PAGE sample preparation kit (Invitrogen) in 1% digitonin according to the manufacturer's instructions. Lysates were run on native PAGE 4–16% Bis-Tris gel and transferred on a PVDF membrane (Invitrogen).

Lentiviral vector production and transduction. HEK293T cells were transfected with pCMVDR8.74 and pMD2.G plasmids, and with the puromycin-selectable lentiviral vector pTRIPZ that contained the open reading frame of respective Flag-mmSTING or Flag-hsSTING constructs, using the calcium phosphate precipitation method. The supernatant that contained lentiviral particles was collected at 48 h and 72 h, pooled and then concentrated by ultracentrifugation. Wild-type mouse embryonic fibroblasts and HEK293T cells were transduced with the lentiviral vectors by adding 5 μ l of concentrated stock directly to the culture medium.

Immunostaining. Mouse embryonic fibroblasts expressing Flag-mmSTING were seeded on coverslips, fixed with 2% (v/v) paraformaldehyde for 10 min, permeabilized for 5 min in 0.1% (v/v) Triton X-100 and blocked with 2% BSA in PBS for 20 min at room temperature. Samples were incubated with the primary antibodies for 3 h (anti-Flag M2 (F1804, Sigma) and GM130 (clone 35, BD Pharmingen,

Alexa Fluor 647). After washing, samples were incubated with secondary antibody (donkey anti-mouse Alexa Fluor 568 (Thermo Fisher)) containing 5 μ g 6-diamino-2-phenylindole (DAPI) for 1 h at room temperature. Coverslips were mounted with Dak (fluorescent mounting medium). Images were acquired by a wide-field fluorescence microscope (Zeiss AxioPlan) and processed in ImageJ software. Confocal sections were obtained with a confocal laser scanning microscope (Zeiss LSM710).

Intact mass measurements for mmSTING. Expression of Flag-mmSTING or Flag-mmSTING(C91S) expression was induced in HEK293T cells overnight by doxycycline. Cells were treated with or without C-178 or C-176 (1 μ M) for 30 min and lysed in lysis buffer (20 mM Hepes, 150 mM NaCl, 10% glycerin and 1% DDM). Flag-STING was immunoprecipitated using anti-Flag M2 affinity gel agarose gel (Sigma) for 2 h at 4°C. Precipitated proteins were eluted using the Flag peptide according to the manufacturer's instructions. Protein mass spectrometry was performed on a Shimadzu MS2020 connected to a Nexerra UHPLC system, equipped with a Waters ACQUITY UPLC BEH C4 1.7- μ m, 2.1 \times 50-mm column. Buffer A was 0.05% formic acid in water, and buffer B was 0.05% formic acid in acetonitrile. The analytical gradient was from 10% to 90% buffer B within 6.0 min with 0.75 ml min⁻¹ flow. Mass spectra were collected from 300–2,000 Da and the spectra were deconvoluted using the software MagTran.

Intact mass measurements for hsSTING and top-down analysis using LC-MS/MS. Proteins were prepared as described in 'Intact mass measurements for mmSTING'. For both intact mass measurement (LC-MS) and top-down analysis (liquid chromatography with high-energy collisional-induced dissociation tandem mass spectrometry, LC-HCD-MS/MS), samples were separated onto column Acquity UPLC Protein BEH C4 (300 \AA , 1.7 μ m, 1 \times 150 mm, Waters) using a Dionex Ultimate 3000 analytical RSLC system (Dionex) coupled to a HESI source (Thermo Fisher Scientific). The separation was performed with a flow rate of 90 μ l min⁻¹ by applying a gradient of solvent B from 15 to 45% in 2 min, then from 45 to 60% within 10 min, followed by column washing and re-equilibration steps. Solvent A was water with 0.1% formic acid, and solvent B was acetonitrile with 0.1% formic acid. Eluting proteoforms were analysed on a high-resolution QExactive HF-Orbitrap-FT-MS benchtop instrument (Thermo Fisher Scientific). For intact mass measurements, MS1 scans were performed in protein mode with 15,000 resolution and averaging 10 micro-scans. Top-down analysis for localization of the compound binding site was performed in PRM mode, isolating specie at 947.5 and 941.6 m/z, with a 300- $\text{\textit{Th}}$ isolation window, 120,000 resolution and averaging 10 micro-scans. HCD was used as fragmentation method with normalized collision energy of 10%, 15% and 19%. Intact mass measurement data were analysed with Protein Deconvolution (Thermo Fisher Scientific) using Respect algorithm with 99% noise rejection confidence and 20 p.p.m. accuracy of average mass identification. Top-down data were deconvoluted using MASH Suite software (GE research group, University of Wisconsin). Data obtained with 3 different NCE values were combined together to create a fragmentation map with assigned *b*- and *y*-fragments using ProSight Lite software (Kelleher research group, Northwestern University) with 20 p.p.m. mass tolerance.

Chemical synthesis. All chemical reagents and anhydrous solvents for synthesis were purchased from commercial suppliers (Sigma-Aldrich, Fluka and Acros) and were used without further purification or distillation. C-176 and C-178 were either purchased (Vitas-M laboratory STK016322 and ChemBridge 5747493) or synthesized. The composition of mixed solvents is given by the volume ratio (v/v). ¹H and ¹³C nuclear magnetic resonance spectra were recorded on a Bruker DPX 400 (400 MHz for ¹H, 100 MHz for ¹³C, respectively) or Bruker AVANCE III 400 Nanobay (400 MHz for ¹H, 100 MHz for ¹³C, respectively) with chemical shifts (δ) reported in p.p.m. relative to the solvent residual signals of DMSO-d6 (2.50 p.p.m. for ¹H, 39.52 p.p.m. for ¹³C). Coupling constants are reported in Hz. LC-MS was performed on a Shimadzu MS2020 connected to a Nexerra UHPLC system equipped with a Waters ACQUITY UPLC BEH C18 1.7- μ m, 2.1 \times 50-mm column. Buffer A was 0.05% HCOOH in H₂O, and buffer B was 0.05% HCOOH in acetonitrile. Analytical gradient was from 10% to 90% buffer B within 6.0 min with flow of 0.5 ml min⁻¹. For details of the chemical synthesis, see Supplementary Information.

Statistical analysis. Prism software (Graphpad Software) was used to perform statistical tests and to generate graphs. Data are presented as mean \pm s.e.m. or s.d. and *P* values were calculated as detailed in the corresponding legends. No statistical methods were used to predetermine sample sizes. Sample sizes were instead determined on the basis of previous experimental experience or based on general practices in the field. Replicates are biological replicates. For the *in vivo* studies, mice were randomly allocated to groups. For histological analysis of tissues, experimenters were blinded to the experimental conditions.

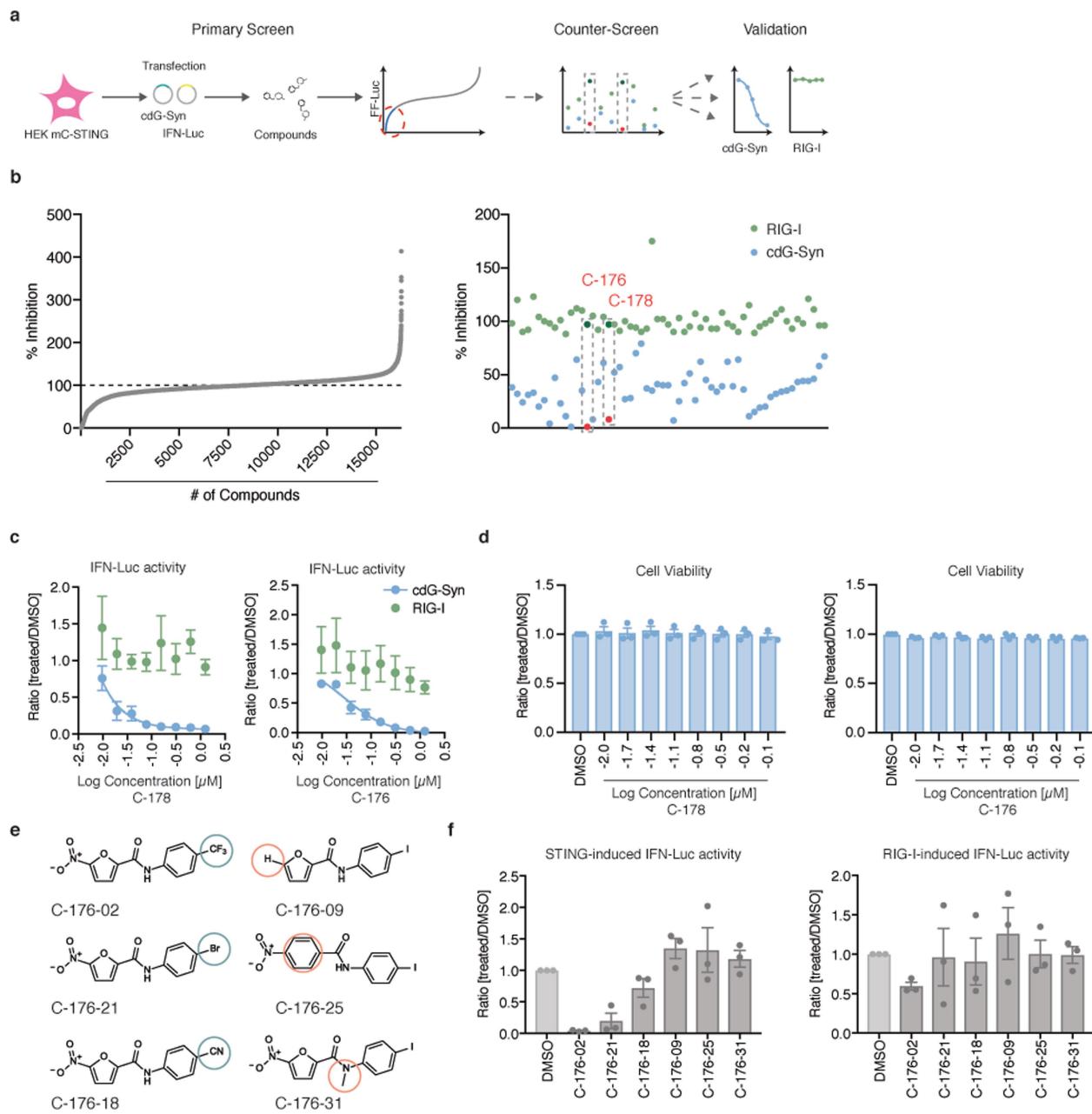
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. RNA sequencing data have been deposited in the Gene Expression Omnibus (GEO) under the accession code GSE113933. Full scans for

all western blots and in-gel fluorescence images are provided. Source Data for animal experiments from Figs 4, 5g–i and Extended Data Figs. 6b–g, 7b–d and 10c are shown in Supplementary Table 1. All other data are available from the corresponding author on reasonable request.

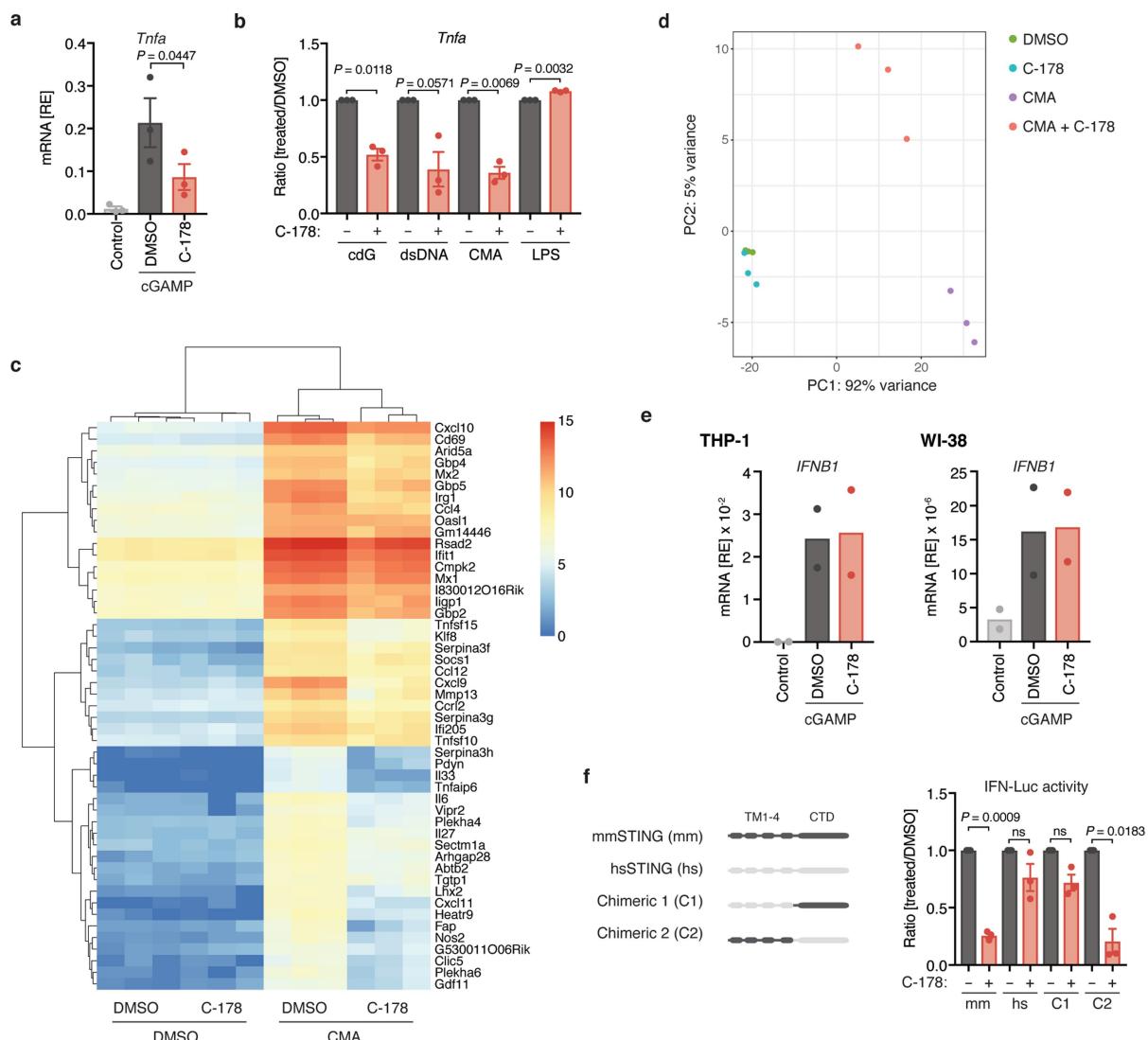
30. Ablasser, A. et al. Cell intrinsic immunity spreads to bystander cells via the intercellular transfer of cGAMP. *Nature* **503**, 530–534 (2013).

31. Morita, M. et al. Gene-targeted mice lacking the Trex1 (DNase III) 3'→5' DNA exonuclease develop inflammatory myocarditis. *Mol. Cell. Biol.* **24**, 6719–6727 (2004).
32. Peschke, K. et al. Loss of Trex1 in dendritic cells is sufficient to trigger systemic autoimmunity. *J. Immunol.* **197**, 2157–2166 (2016).
33. Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C. & Deplancke, B. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* **33**, 3123–3125 (2017).



Extended Data Fig. 1 | A chemical screen identifies small-molecule inhibitors of STING. **a**, Screening workflow. HEK mC-STING, HEK293T cells expressing mCherry-STING. **b**, Left, summary of the primary screen. Mean of normalized values for IFN β luciferase activity in cells treated with compound versus cells treated with DMSO. Right, validation of selected candidates from the primary screen. Normalized values for IFN β luciferase activity (light blue) induced by coexpression of cyclic di-GMP synthase (cdG-Syn) and STING are shown, in comparison to activity triggered

by RIG-I (green). **c, d**, HEK293T cells expressing mCherry-STING were transfected with plasmids that encoded either cdG-Syn or RIG-I, as well as an IFN β luciferase reporter, and then treated with C-178 or C-176 (1.25 μ M–0.01 μ M), after which IFN β luciferase activity (**c**) or cell viability (using the CellTiter-Blue assay) (**d**) were measured. **e, f**, Chemical structures of derivatives of C-176 (**e**) and their effect (at a concentration of 0.5 μ M) on IFN β luciferase reporter activity triggered by mmSTING or RIG-I (**f**). Mean \pm s.e.m. of $n = 3$ experiments (**c, d, f**).

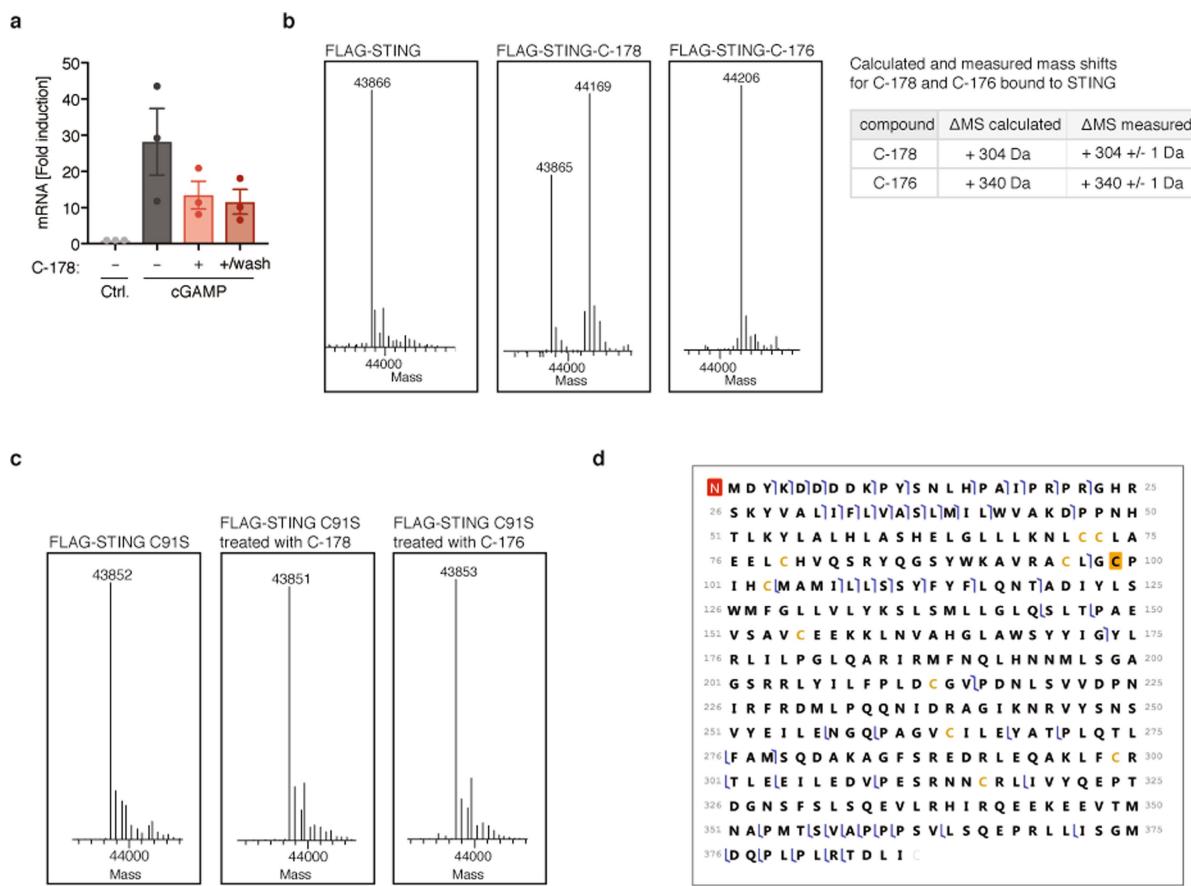


Extended Data Fig. 2 | Activity of C-178 against STING in distinct cells.

a, b, mRNA expression levels of *Tnfa* in BMDMs activated with cGAMP, cyclic di-GMP (cdG), dsDNA, CMA or LPS for 5 h, after pretreatment for 1 h with C-178 (0.5 μ M) or DMSO ($n = 3$ biological replicates).

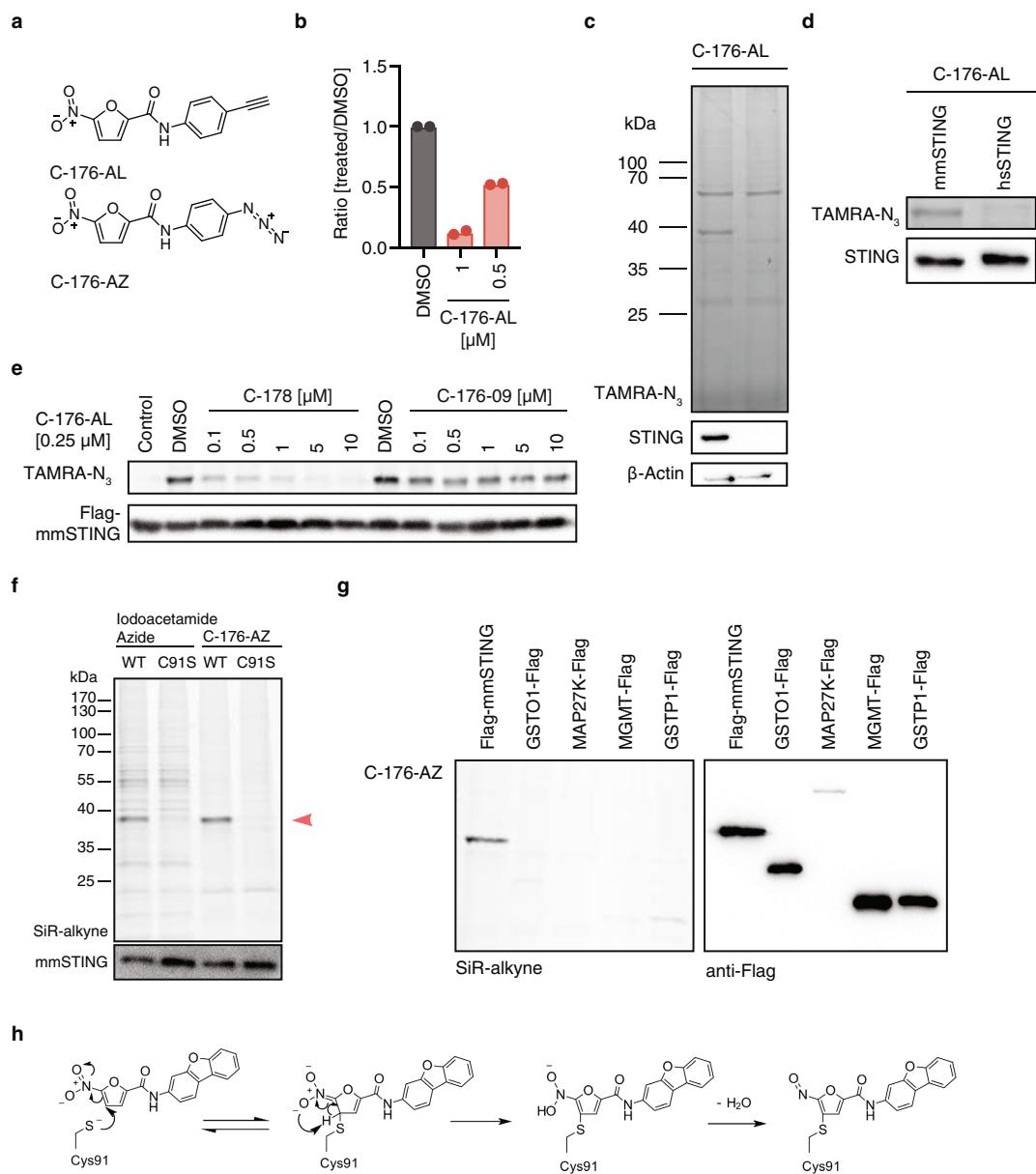
c, Heat map of RNA sequencing analysis of BMDMs treated with DMSO or CMA, in the presence or absence of C-178. The top 50 upregulated genes in cells treated with DMSO versus cells treated with CMA, in the absence of C-178, are shown for all conditions. **d**, Scatter plot of the dimensions PC1 versus PC2. **e**, Levels of expression of *IFNB1* mRNA in

THP-1 and WI-38 cells pretreated with C-178 (0.25 μ M) or DMSO and stimulated with cGAMP for 3 h ($n = 2$ biological replicates). **f**, HEK293T cells were transfected with the construct chimeric 1 (C1: hsSTING (amino acids 1–138)–mmSTING (amino acids 138–378)) or the construct chimeric 2 (C2: mmSTING (amino acids 1–137)–hsSTING (amino acids 139–379)), together with cdG-Syn and IFN β luciferase reporter, and treated with C-178 (0.5 μ M) ($n = 3$ biological replicates). Data are mean or mean \pm s.e.m. P values were determined by two-tailed *t*-test.



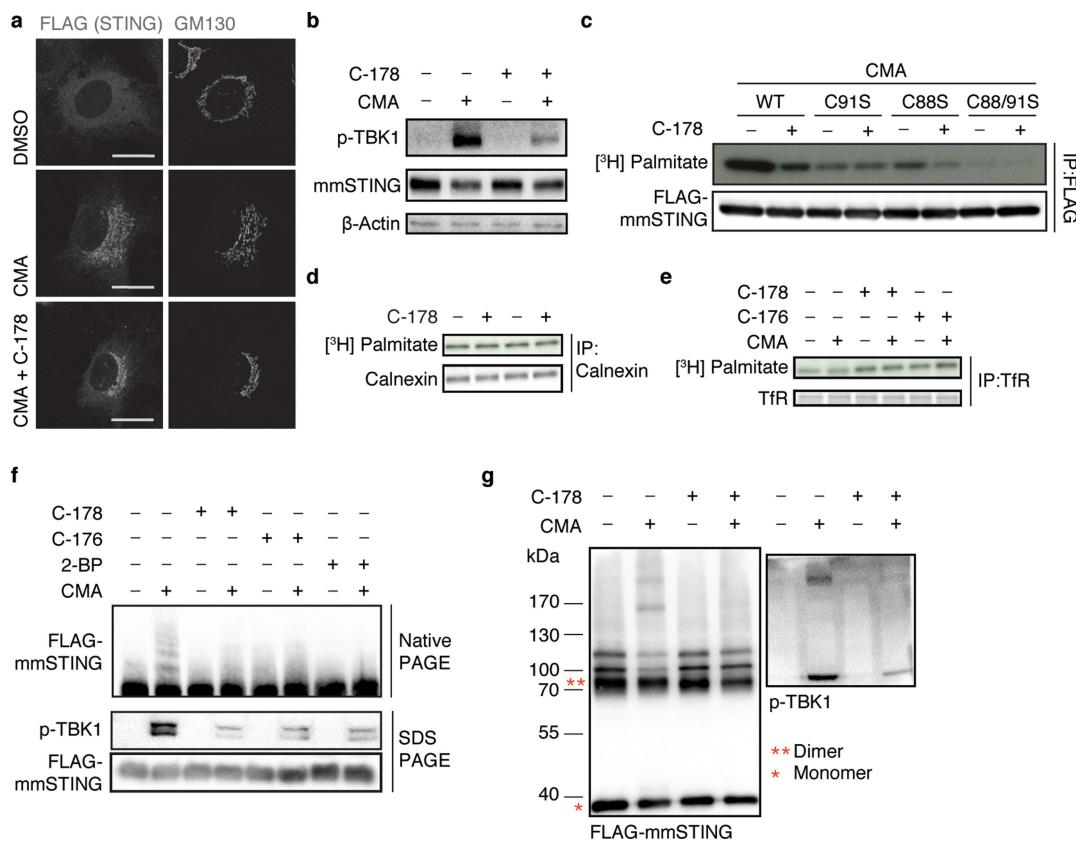
Extended Data Fig. 3 | C-178 and C-176 covalently bind to Cys91.
a, BMDMs were treated with C-178 (0.25 μ M) for 1 h, washed or left untreated and after 1 h were stimulated with cGAMP for 3 h. Levels of *Ifnb1* mRNA were assessed by RT-qPCR. Data are mean \pm s.e.m. ($n=3$ biological replicates). **b-d**, HEK293T cells expressing indicated Flag-mmSTING constructs were exposed to C-178 or C-176 (1 μ M), lysed and Flag-mmSTING was immunoprecipitated. The eluted protein was analysed by intact mass spectrometry (LC-MS). **b**, Deconvoluted electrospray ionization mass spectrum (left), and expected and measured mass shifts after covalent binding of C-176 and C-178 (right) are shown.

c, Deconvoluted electrospray ionization mass spectrum for Flag-mmSTING(C91S). **d**, Fragmentation map of Flag-mmSTING-C-176 analysed by top-down analysis, using LC-HCD-MS/MS (an additional 8 residues are due to N-terminal Flag). Data that were obtained with three different NCE values are combined to create a fragmentation map with assigned *b*- and *y*-fragments. Achieved sequence coverage is 15% with 20 p.p.m. mass accuracy for fragment assignment. One representative of $n=2$ independent experiments is shown (**b**, **c**). For electrospray ionization spectra see Supplementary Information.



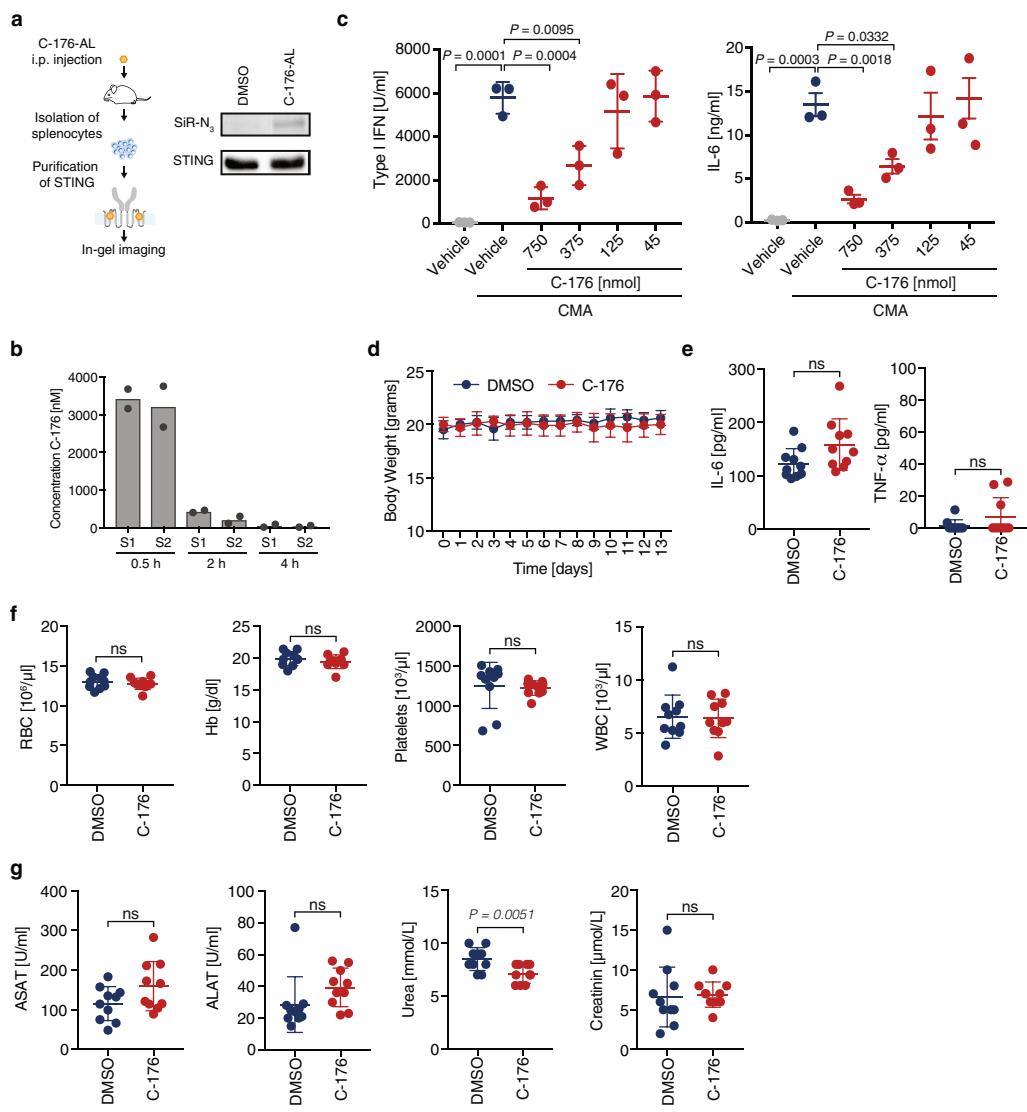
Extended Data Fig. 4 | Gel-based analysis of compound interactions using clickable probes. a, Structures of C-176-AL and C-176-AZ. **b**, HEK293T cells transfected with Flag-mmSTING and an IFN β luciferase reporter were treated with C-176-AL, and luciferase activity was then assessed. Data are mean of $n = 2$. **c**, Labelling events of wild-type HEK293T cells and HEK293T with Flag-mmSTING incubated with C-176-AL (0.25 μ M). **d**, Distinct labelling of mmSTING and hsSTING by C-176-AL (0.25 μ M). **e**, Concentration-dependent competitor blockage of C-176-AL (0.25 μ M) against C-178 and C-176-09 in HEK293T cells that

express Flag-mmSTING. **f**, Labelling events of HEK293T cells that express Flag-mmSTING (wild-type mmSTING or mmSTING(C91S)) when exposed to iodoacetamide azide or C-176-AZ (both at 0.25 μ M). **g**, HEK293T cells expressing Flag-mmSTING, GSTO1-Flag, MAP27K-Flag, MGMT-Flag or GSTP1-Flag were treated with C-176-AZ (0.25 μ M). For in-gel fluorescence imaging, TAMRA azide (**c–e**) or SiR alkyne (**f, g**) was used. Immunoblots against Flag or β -actin are shown and data are representative of $n = 3$ independent experiments (**c–g**). **h**, Proposed reaction mechanism. For gel source data, see Supplementary Fig. 1.



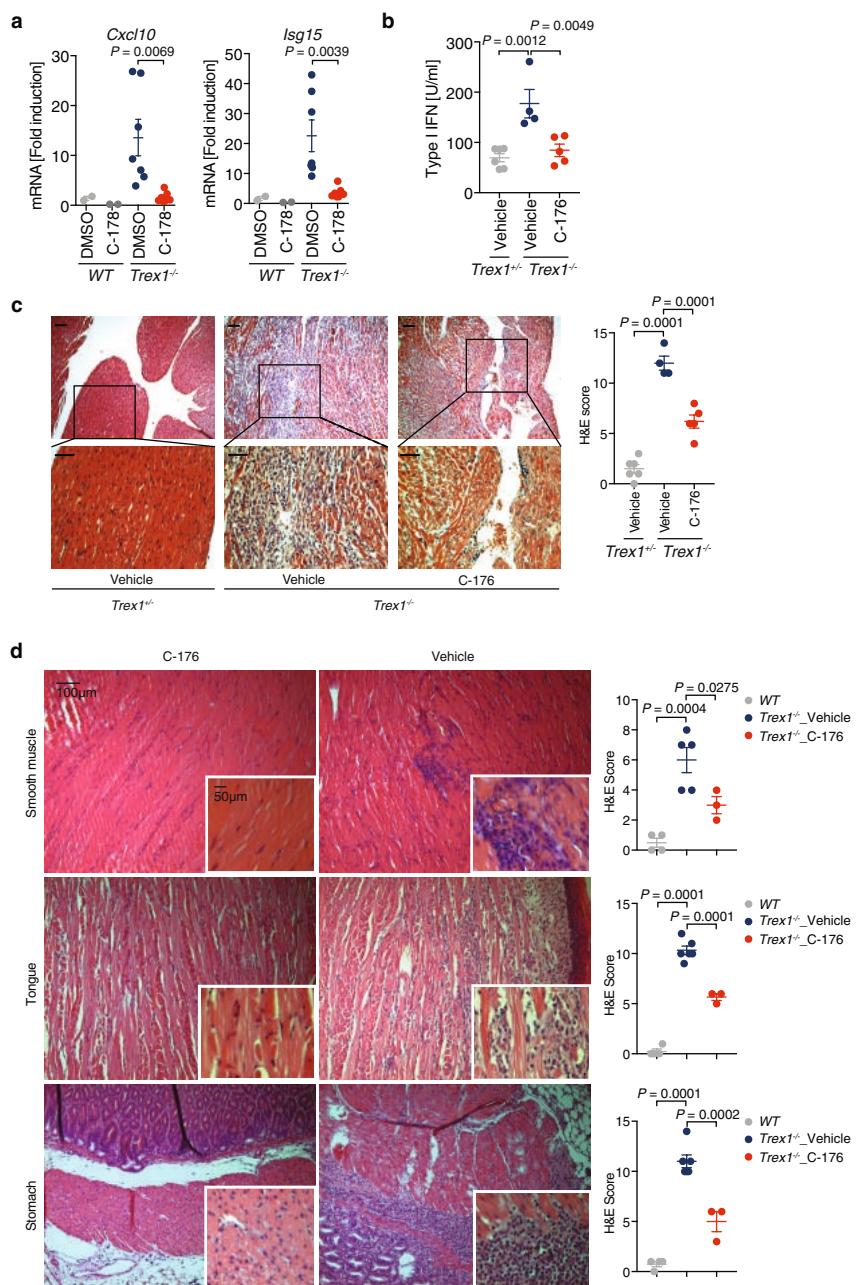
Extended Data Fig. 5 | C-178 and C-176 block activation-induced palmitoylation of STING. **a**, Single-channel images of Flag and GM130 stainings of mouse embryonic fibroblasts that express Flag-mmSTING (control corresponding to images shown in Fig. 3), treated with CMA or DMSO. Scale bar, 20 μ m. **b**, Protein quantification of p-TBK1, Flag-mmSTING and β -actin by immunoblot of mouse embryonic fibroblasts from **a**. **c**, [³H]-palmitate labelling of HEK293T cells that express the indicated Flag-mmSTING constructs, and which were treated with C-178, C176 (1 μ M) or DMSO. **d**, **e**, [³H]-palmitate labelling of immunoprecipitated endogenous calnexin or transferrin receptor from

HEK293T cells treated as indicated (compounds at 1 μ M). **f**, HEK293T cells that express Flag-mmSTING were treated with C-178 (1 μ M) or 2-bromopalmitate (2-BP) (50 μ M) and stimulated with CMA for 1.5 h. Analysis of indicated proteins was performed by native PAGE or SDS-PAGE. **g**, Crosslinked lysates (DSS, 1 mM) of HEK293T cells that express Flag-mmSTING treated with C-178 (1 μ M) and stimulated with CMA (2 h) were analysed by SDS-PAGE and immunoblotted for STING and p-TBK1. One representative of $n = 3$ (**a–c**, **e–g**) or $n = 2$ (**d**) independent experiments. For gel source data, see Supplementary Fig. 1.



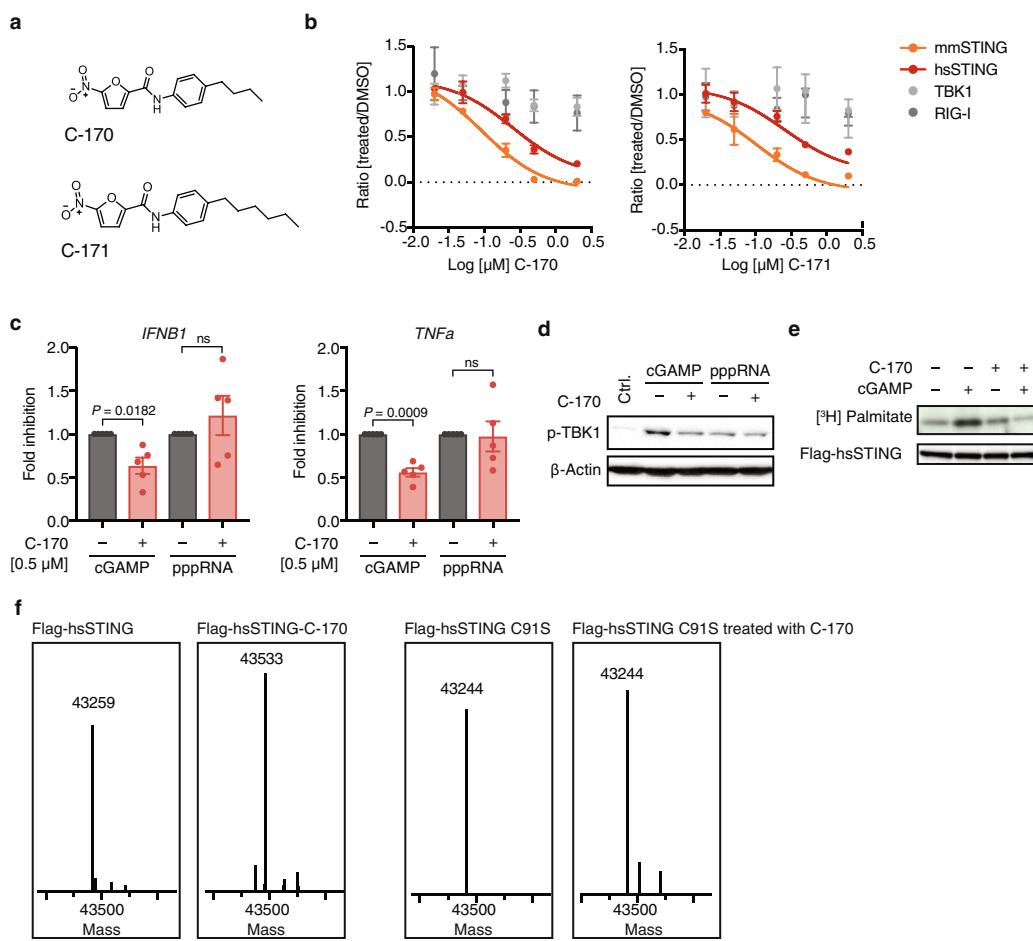
Extended Data Fig. 6 | In vivo effects of C-176 in wild-type mice.
a, Labelling of endogenous STING immunoprecipitated from splenocytes of mice treated with C-176-AL, visualized by in-gel fluorescence. One representative of $n = 2$. **b**, Plasma concentration profiles of C-176 after a single-dose intraperitoneal injection into wild-type mice ($n = 2$ mice per condition). Data are mean of technical replicates. **c**, Serum levels of type I IFNs and IL-6 from wild-type mice pretreated with C-176 or vehicle 4 h after injection with CMA ($n = 3$ mice per condition). **d**, Body weight of

wild-type mice during two weeks of daily DMSO and C-176 injection. **e–g**, Mice from **d** were euthanized and blood samples were collected for measuring plasma levels of TNF α and IL-6 (**e**), blood cell counts (**f**), and liver and kidney parameters (**g**). Data are mean \pm s.d. of $n = 10$ mice per condition (**e–g**). P values were calculated by one-way ANOVA (**c**) or two-tailed t -test (**e–g**). ns, not significant. For gel source data, see Supplementary Fig. 1. For source data, see Supplementary Table 1.



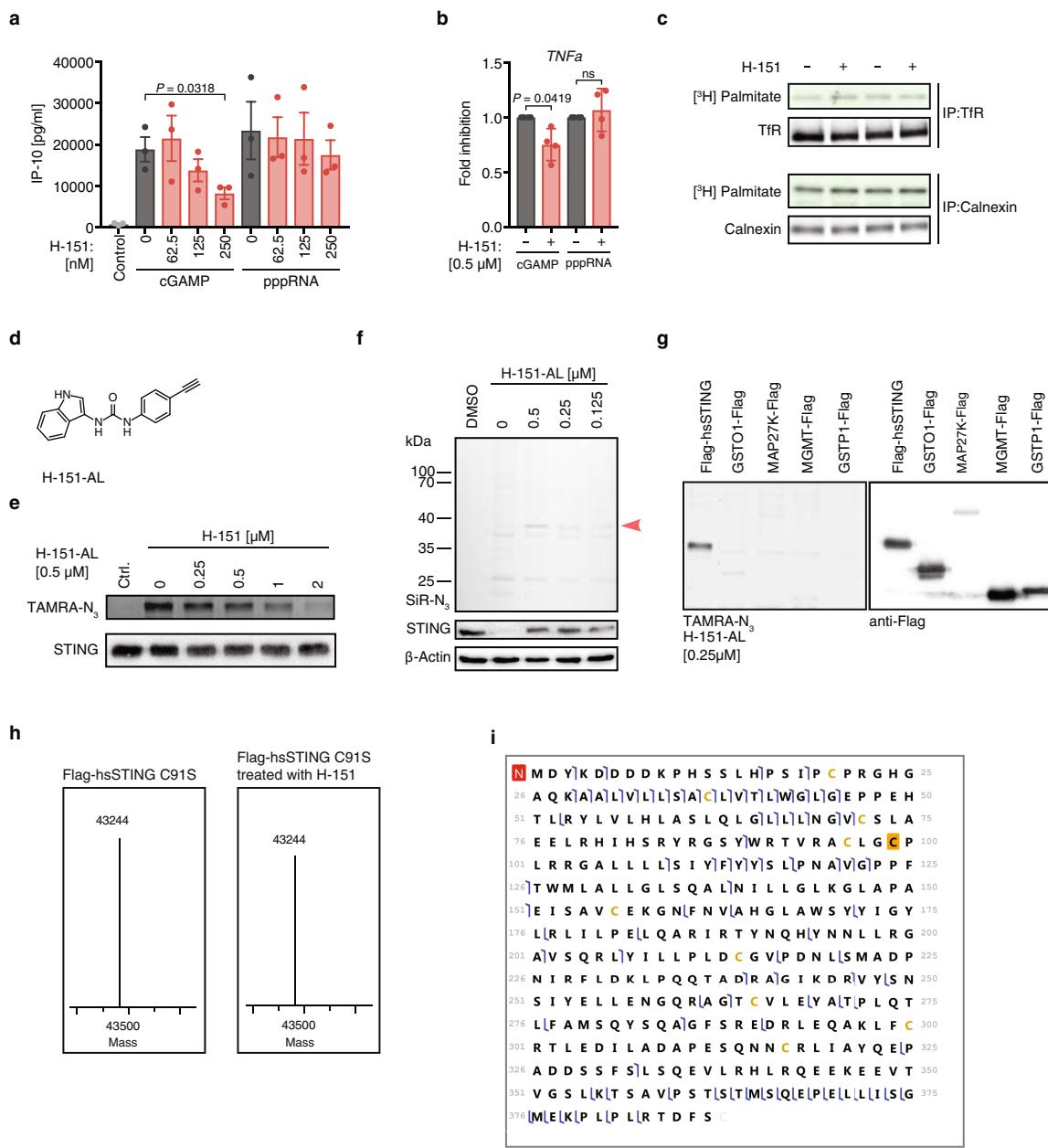
Extended Data Fig. 7 | Activity of C-178 and C-176 in TREX1-deficient cells and mice. **a**, Wild-type ($n=2$) or $Trex1^{-/-}$ ($n=7$) mouse embryonic fibroblasts were treated with DMSO or C-178 (2 μ M) overnight. mRNA levels of *Isg15* and *Cxcl10* were measured by RT-qPCR. **b, c**, Control mice (wild type or $Trex1^{+/+}$ (each $n=3$)) were treated with vehicle, and $Trex1^{-/-}$ mice were treated with C-176 ($n=5$ mice) or vehicle ($n=4$ mice) for 11 days. Serum type I IFN levels (**b**) were measured and histological

analysis of the heart (**c**) was performed. Scale bars, 50 μ m. **d**, Histological analysis of distinct organs from wild-type (all $n=4$) or $Trex1^{-/-}$ mice treated for 3 months. Smooth muscle, $n=3$ (C-176) or 5 (vehicle); tongue and stomach, $n=3$ (C-176) or 6 (vehicle). Representative histological images are shown. Data are mean \pm s.e.m. P values were calculated using two-tailed *t*-test (**a**) or one-way ANOVA (**b-d**). For source data, see Supplementary Table 1.



Extended Data Fig. 8 | C-170 and C-171 antagonize hsSTING.
a, Structure of C-170 and C-171. **b**, IFN β luciferase reporter measurements from HEK293T cells transfected with indicated constructs (C-170 and C-171, 0.02–2 μ M) ($n = 3$). **c**, **d**, THP-1 cells were pretreated with C-170 (0.5 μ M) and stimulated with cGAMP or triphosphate RNA. *IFNB1* and *TNF α* mRNA levels were assessed by RT-qPCR ($n = 5$) (**c**), and p-TBK1 was determined by immunoblot (**d**). **e**, [3 H]-palmitate labelling of HEK293T cells that express Flag-hsSTING, treated with C-170 (1 μ M) or DMSO. **f**, HEK293T cells that express Flag-hsSTING (wild-type

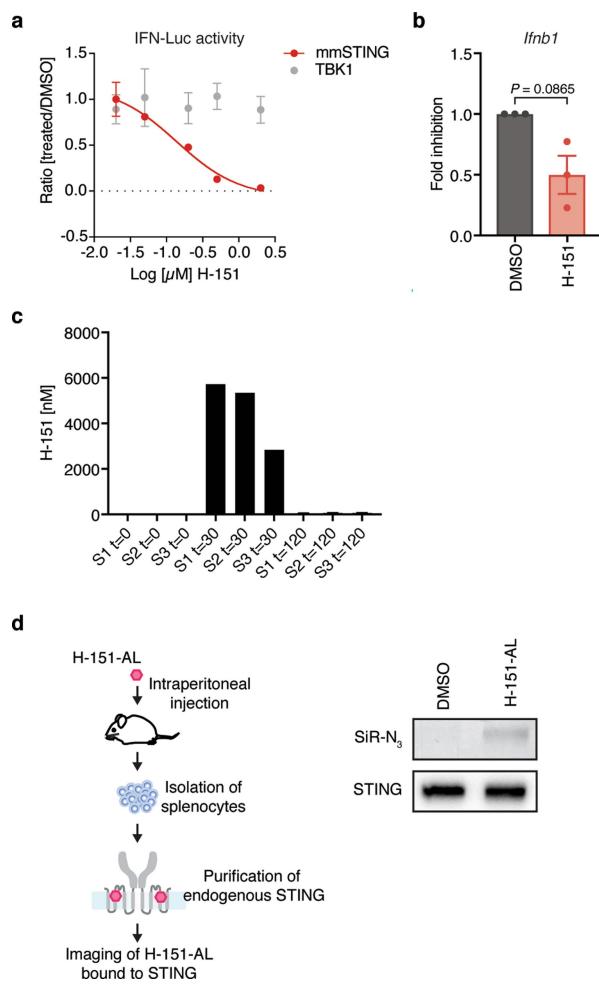
hsSTING or hsSTING(C91S)) were treated with C-170 (1 μ M), lysed and Flag-hsSTING was analysed by intact mass measurement (LC-MS). Deconvoluted electrospray ionization mass spectrum showing intact mass indicated hsSTING constructs and treatments, are shown. Data are mean \pm s.e.m. *P* values were calculated using two-tailed *t*-test. NS, not significant. One representative of $n = 3$ (**d**) and $n = 2$ (**e-f**) independent experiments is shown. For gel source data, see Supplementary Fig. 1. For electrospray ionization spectra see Supplementary Information.



Extended Data Fig. 9 | Mechanism of STING inhibition by H-151.

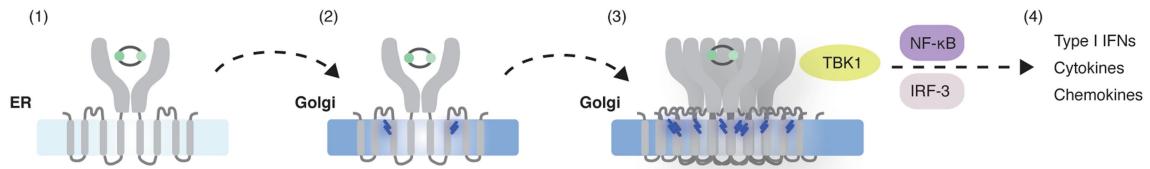
a, THP-1 cells that had been pretreated with H-151 or DMSO were stimulated with cGAMP or triphosphate RNA, or left unstimulated. IP-10 production was quantified by enzyme-linked immunosorbent assay after overnight incubation ($n = 3$). **b**, Levels of TNF mRNA assessed by RT-qPCR in THP-1 cells that had been pretreated with H-151 and stimulated with cGAMP or triphosphate RNA ($n = 4$). **c**, $[^3\text{H}]$ -palmitate labelling of immunoprecipitated endogenous calnexin or transferrin receptor from HEK293T cells that had been treated with H-151 (1 μM). **d**, Structure of H-151-AL. **e**, Competition assay of H-151-AL with H-151 in HEK293T cells that express Flag-hsSTING. Flag-hsSTING labelled with H-151-AL was visualized by in-gel fluorescence. **f**, HEK293T cells or HEK293T cells that express Flag-hsSTING were treated with H-151-AL, lysed and clicked to a SiR azide. Whole-cell lysates were analysed by in-gel fluorescence and by immunoblot. **g**, HEK293T cells that express Flag-

hsSTING, GSTO1-Flag, MAP27K-Flag, MGMT-Flag or GSTP1-Flag were exposed to H-151-AL, clicked to TAMRA azide and visualized by in-gel fluorescence or immunoblot. **h**, Deconvoluted electrospray ionization mass spectrum showing intact mass of Flag-hsSTING(C91S) with or without H-151. **i**, Flag-hsSTING was purified from HEK293T cells that had been pretreated with H-151 (1 μM) and analysed by top-down analysis using LC-HCD-MS/MS (with additional 8 residues due to N-terminal Flag). Data that were obtained with three different NCE values are combined to create fragmentation map with assigned *b*- and *y*-fragments. Achieved sequence coverage is 23% with 20 p.p.m. mass accuracy for fragment assignment. Data are shown as mean \pm s.e.m. *P* values were calculated by two-tailed *t*-test. NS, not significant. One of $n = 2$ experiments is shown (**c**, **e–h**). For gel source data, see Supplementary Fig. 1. For electrospray ionization spectra, see Supplementary Information.

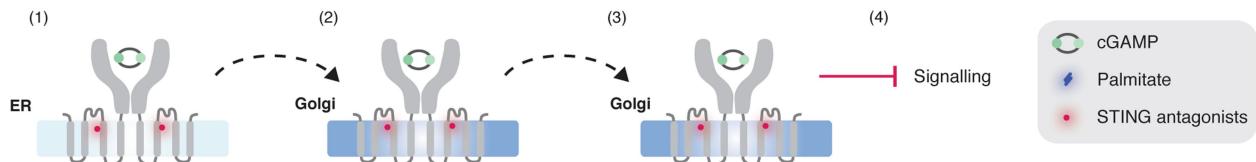

Extended Data Fig. 10 | Activity of H-151 against mmSTING.

a, HEK293T cells were transfected with plasmids encoding mmSTING in combination with cdG-Syn or were transfected with a plasmid for TBK1, and an IFN β luciferase reporter, and then treated with H-151 (concentration 0.02–2 μ M). Reporter activity was measured after overnight incubation. Data are mean \pm s.e.m. ($n = 3$; nonlinear regression analysis). Experiments were performed together with those that generated the data displayed in Fig. 5b. **b**, BMDMs were pretreated with H-151 (0.5 μ M) and levels of *Ifnb1* mRNA expression induced by cGAMP were assessed by RT-qPCR. Data are mean \pm s.e.m. ($n = 3$); P value was calculated by two-tailed *t*-test. **c**, Mean plasma concentration profiles of H-151 following a single-dose intraperitoneal injection into wild-type mice ($n = 3$ mice per time point). **d**, Schematic, and in vivo detection of H-176-AL binding to mmSTING. Visualization by in-gel fluorescence was performed using SiR azide. One representative of $n = 2$. For gel source data, see Supplementary Fig. 1. For source data, see Supplementary Table 1.

Intact STING signalling



Small-molecule mediated inhibition



Extended Data Fig. 11 | Mechanism of action of the identified STING antagonists. In the absence of inhibition, ligand binding triggers the translocation of STING to the Golgi^{1,25}, where palmitoylation occurs at cytoplasmic proximal cysteine residues (Cys88 and Cys91)²⁶. In turn, this post-translational modification facilitates the multimerization of

STING to create a platform—possibly at the lipid raft domain at the trans-Golgi network²⁶—for the recruitment of TBK1, and thereby enables the initiation of downstream signalling. Through covalent interaction with Cys91, the compounds we describe here block the palmitoylation of STING and retain the protein in a signalling-incompetent state.

Mechanism for remodelling of the cell cycle checkpoint protein MAD2 by the ATPase TRIP13

Claudio Alfieri¹, Leifu Chang^{1,2} & David Barford^{1*}

The maintenance of genome stability during mitosis is coordinated by the spindle assembly checkpoint (SAC) through its effector the mitotic checkpoint complex (MCC), an inhibitor of the anaphase-promoting complex (APC/C, also known as the cyclosome)^{1,2}. Unattached kinetochores control MCC assembly by catalysing a change in the topology of the β -sheet of MAD2 (an MCC subunit), thereby generating the active closed MAD2 (C-MAD2) conformer^{3–5}. Disassembly of free MCC, which is required for SAC inactivation and chromosome segregation, is an ATP-dependent process driven by the AAA+ ATPase TRIP13. In combination with p31^{comet}, an SAC antagonist⁶, TRIP13 remodels C-MAD2 into inactive open MAD2 (O-MAD2)^{7–10}. Here, we present a mechanism that explains how TRIP13–p31^{comet} disassembles the MCC. Cryo-electron microscopy structures of the TRIP13–p31^{comet}–C-MAD2–CDC20 complex reveal that p31^{comet} recruits C-MAD2 to a defined site on the TRIP13 hexameric ring, positioning the N terminus of C-MAD2 (MAD2^{NT}) to insert into the axial pore of TRIP13 and distorting the TRIP13 ring to initiate remodelling. Molecular modelling suggests that by gripping MAD2^{NT} within its axial pore, TRIP13 couples sequential ATP-driven translocation of its hexameric ring along MAD2^{NT} to push upwards on, and simultaneously rotate, the globular domains of the p31^{comet}–C-MAD2 complex. This unwinds a region of the α A helix of C-MAD2 that is required to stabilize the C-MAD2 β -sheet, thus destabilizing C-MAD2 in favour of O-MAD2 and dissociating MAD2 from p31^{comet}. Our study provides insights into how specific substrates are recruited to AAA+ ATPases through adaptor proteins and suggests a model of how translocation through the axial pore of AAA+ ATPases is coupled to protein remodelling.

To investigate disassembly of the MCC catalysed by the joint action of p31^{comet} and TRIP13, we incubated p31^{comet} and TRIP13 with ATP and either free MCC or MCC in complex with the APC/C (APC/C–MCC). p31^{comet} and TRIP13 catalysed the formation of O-MAD2 from the MCC much more effectively than from APC/C–MCC (Extended Data Figs. 1a (lanes 5–8), b, 2a). Release of O-MAD2 from APC/C–MCC did not require the APC/C subunit APC15 (Extended Data Fig. 1a (lanes 1, 2 and 5, 6), consistent with TRIP13–p31^{comet} and the APC/C mediating independent SAC silencing pathways^{8,9,11–16}. The TRIP13 catalytic mutant (TRIP13(E253Q))¹⁷ is defective in generating O-MAD2 (Extended Data Fig. 1a (lanes 3, 4, 9, 10), b). When incubated with the MCC and p31^{comet}, a TRIP13(E253Q) hexamer formed a complex with p31^{comet}, CDC20 and MAD2 in a 1:1:1 stoichiometry, consistent with previous findings^{7,18} (Extended Data Fig. 1c). p31^{comet}, but not TRIP13, interacts with the free MCC and APC/C–MCC (Extended Data Fig. 1b, d). The substrate of the TRIP13 ATPase reaction is the p31^{comet}–C-MAD2–CDC20 complex (referred to hereafter as p31–substrate), which forms a tight complex with TRIP13(E253Q) (Extended Data Fig. 1e).

We then determined the cryo-EM structure of the TRIP13(E253Q)–p31–substrate complex with ATP γ S (Extended Data Fig. 1e). Refinement of the electron microscopy data set yielded a 3D reconstruction at 4.5 Å resolution, extending to 4.3 Å for focused refinement of TRIP13 monomers A–D (Extended Data Fig. 3, Extended Data Table 1). 3D classification revealed three structural classes

(Extended Data Fig. 4). Class 1 is the unliganded apo TRIP13 hexamer, and classes 2 and 3 belong to two distinct conformational states of the TRIP13(E253Q)–p31–substrate complex. TRIP13 in class 2 represents the basal state, whereas class 3 presents structural rearrangements in TRIP13, suggestive of an activated state. An atomic model for apo TRIP13 was built into the overall electron microscopy density map, guided by the crystal structure of a TRIP13 monomer¹⁸ (Fig. 1a, Extended Data Fig. 4a, Extended Data Table 1). In the cryo-electron microscopy (cryo-EM) structure, TRIP13 forms a closed hexameric ring featuring a central pore, whereas in the crystal structure adjacent subunits assemble into a helical filament¹⁸ (Fig. 1a, Extended Data Fig. 5a). The TRIP13 cryo-EM structure also differs from the flat closed ring of the *Caenorhabditis elegans* TRIP13 orthologue PCH2¹⁷ (Fig. 1a, Extended Data Fig. 5a).

Our cryo-EM structure shows that with ATP γ S the TRIP13 ring adopts a compact conformation with a convex top surface and concave lower surface (Fig. 1a, Extended Data Fig. 5a). Monomers A–E form a right-handed spiral. Monomer F, which is structurally distinct from its counterparts owing to the separation of its large and small AAA+ domains (although less pronounced than that of the open subunits of PCH2¹⁷; Extended Data Fig. 5b), forms a seam in the hexameric ring that bridges monomers A and E (Fig. 1a, Extended Data Fig. 5a, b). Clear cryo-EM density is visible for ATP γ S at monomers A–E (Extended Data Fig. 3e). Pore loop-1, which occludes the central pore in the PCH2 flat ring conformation, rotates upwards (Extended Data Fig. 5c) such that Trp221 and Phe222 of monomer A are exposed on the surface.

To build the basal state TRIP13(E253Q)–p31–substrate model (class 2 of the 3D classification; Extended Data Fig. 4c), the crystal structure of the p31^{comet}–C-MAD2 dimer¹⁹ was docked into the large cryo-EM density feature above the convex face of TRIP13 (Fig. 1b, Extended Data Fig. 3c, d). The p31^{comet}–C-MAD2 dimer undergoes little conformational change except for a structural rearrangement of MAD2^{NT} (Extended Data Fig. 6a, b). The conformation of the TRIP13 ring is essentially identical to that of apo TRIP13 (Extended Data Fig. 4b). The main contact between p31–substrate and TRIP13 involves a conserved basic surface on p31^{comet} opposite the C-MAD2 interface (Figs. 1b, 2a, b and Supplementary Video 1). Strikingly, p31^{comet} positions C-MAD2 so that MAD2^{NT}, which extends from the start of the α A helix, is precisely located to enter the TRIP13 pore (Figs. 1b, 3a, Extended Data Fig. 3d, f). Notably, structural elements of C-MAD2 that are remodelled on conversion to O-MAD2—the β 8'– β 8'' hairpin and safety belt (Extended Data Fig. 6b, c)—form no direct contacts with TRIP13.

TRIP13 interacts with p31^{comet} through a conserved and negatively charged surface on monomers D and E. Monomer D interacts with the safety belt of p31^{comet}, with monomer E contacting conserved basic residues of the extended p31^{comet} α 3–4 loop (Figs. 1b, 2b, Extended Data Fig. 7a–c), consistent with previous results^{17,18,20}. Combined deletion of seven basic residues of the α 3–4 loop severely ablated TRIP13-catalysed remodelling of MAD2, without disrupting complex assembly (Fig. 2c, Extended Data Figs. 2c, 7d).

¹MRC Laboratory of Molecular Biology, Cambridge, UK. ²Present address: Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. *e-mail: dbarford@mrc-lmb.cam.ac.uk

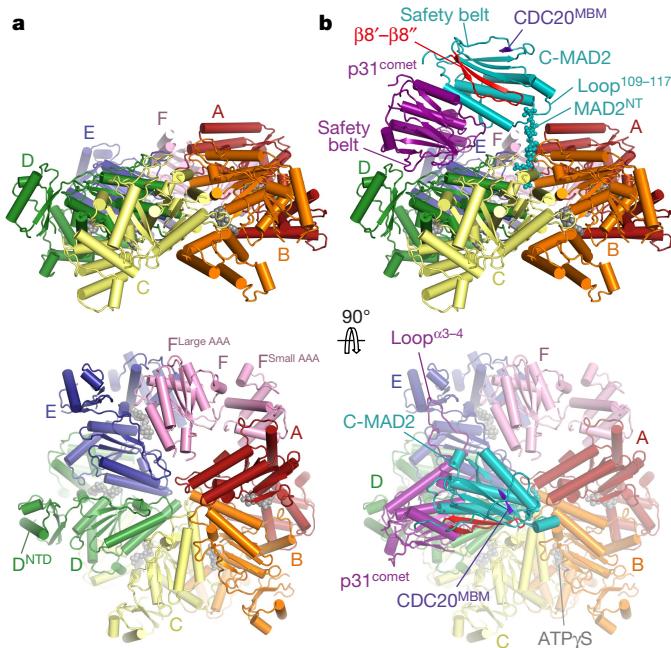


Fig. 1 | Overall structures of the apo and TRIP13-p31-comet-substrate complexes. **a, b**, Side and top views of the cryo-EM structures of TRIP13 in the apo state (a) and in complex with p31^{comet}-C-MAD2-CDC20 (class 2; basal state) (b). Relevant regions of TRIP13, p31^{comet} and C-MAD2 are indicated. F^{Large AAA} and F^{Small AAA}: large and small AAA+ domains, respectively, of monomer F.

In the TRIP13-p31-comet-substrate complex, C-MAD2 interacts tightly with p31^{comet} through the MAD2 dimerization interface^{19,21-23} and entraps the MAD2-binding motif (MBM) of CDC20 under its safety belt (Fig. 1b). No other density for CDC20 was visible (Extended Data Fig. 3c, d). As previously reported¹⁸, the 109–117 loop of C-MAD2 interacts with pore loop-1 of TRIP13 monomer A. MAD2^{NT} adopts a

fully extended conformation with residues 2–9 inserting mid-way into the narrow pore at the centre of TRIP13 (Figs. 1b, 3a, Extended Data Fig. 3d, f and Supplementary Video 1). In this configuration, the conserved TRIP13 pore loop-1 Trp221 and Phe222 of monomers B–D embrace residues 2–7 of MAD2^{NT} (Fig. 3a, Extended Data Fig. 3f). Trp221 and Phe222 are responsible for MAD2 processing¹⁷. Pore loop-1 of monomer E is too low in the TRIP13 spiral to contact MAD2^{NT} (Fig. 3a), whereas that of monomer F is disordered.

Monomer A of TRIP13 also interacts with MAD2^{NT} through the underlying acidic pore loop-2 (Fig. 3a), contributing a key specificity determinant for MAD2^{NT}. The region of MAD2^{NT} that forms intimate contacts with conserved pore loop residues of monomers A–D is hydrophobic and basic (L³QLSR⁷) (Fig. 3a). Although the cryo-EM density for side chains of MAD2^{NT} and pore loops-2 is incomplete, it is indicative of interactions between Arg7 of MAD2^{NT} and Glu269 and Asp272 of pore loop-2 of monomer A, with Leu3 and Leu5 of MAD2^{NT} contacting the Phe222 side chains of pore loops-1 of monomers C and B, respectively.

Notably, the MAD2^{NT} L³QLSR⁷ motif is conserved in MAD2 homologues and other TRIP13 HORMA-domain protein substrates¹⁸ (Extended Data Fig. 8c). To test the importance of the MAD2^{NT} LSR sequence, we compared the efficacy of O-MAD2 release from a C-MAD2-CDC20 complex incorporating either wild-type C-MAD2 or a mutant in which LSR was replaced by LEE. Release of O-MAD2 was reduced in the LEE mutant (Fig. 3b (lanes 3–4, 7–8), Extended Data Fig. 2c), showing that a positively charged segment in MAD2^{NT} is important for TRIP13 function. A TRIP13 mutant in which Glu269 and Asp272 were replaced with either Ala or Arg was completely defective in MAD2 remodelling (Fig. 3c, Extended Data Fig. 2c). Removal of the first seven residues of MAD2 severely impaired O-MAD2 release (Fig. 3b (lanes 3–6), Extended Data Fig. 2b), but deleting up to nine residues from the MAD2 N terminus did not affect complex assembly (Extended Data Fig. 7d). These findings agree with a previous study, in which deletion of the first five residues of MAD2 (Δ N5-MAD2) abolished both the ability of TRIP13 to convert C-MAD2 to O-MAD2, and the substrate-dependent stimulation of TRIP13 ATPase activity¹⁸.

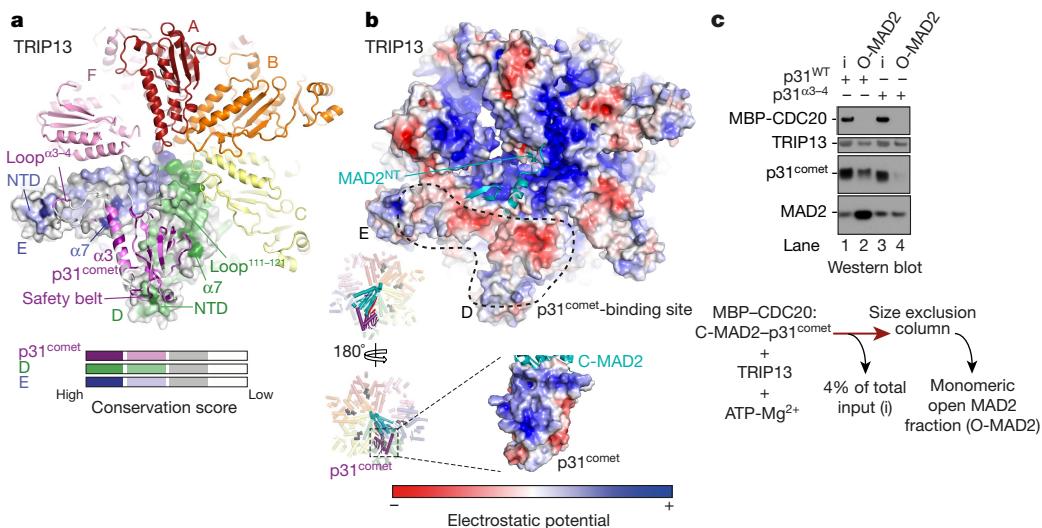


Fig. 2 | Interaction between p31^{comet} and TRIP13. **a**, Surface conservation is displayed for TRIP13 monomers D and E (conservation score colour code indicated below). A portion of p31^{comet} is shown in cartoon representation. **b**, Electrostatic surface representation of the TRIP13-p31^{comet} interacting surfaces (surface potential at ± 5 kT e⁻¹, colour code displayed below). Top, electrostatic surface of TRIP13 with a portion of C-MAD2 shown in cartoon representation. Monomers D and E form a continuous acidic patch that is unique to this monomer pairing owing to the conformation of loop 111–121 of monomer E. Bottom, electrostatic surface of p31^{comet} as viewed by TRIP13. Cartoons to the left

provide an overall perspective. The surface area buried at the p31^{comet}-TRIP13 interface is 2,392 Å². **c**, Deletion of conserved basic residues of the p31^{comet} α 3–4 loop (Extended Data Fig. 7b) severely disrupts TRIP13-p31^{comet}-catalysed remodelling of MAD2. Bottom, experimental design for CDC20-C-MAD2 complex disassembly reaction. Top, western blot showing input (i) and size exclusion fractions corresponding to monomeric O-MAD2 (lanes 2 and 4 and Extended Data Fig. 2c). The experiment shown in c was performed in triplicate with similar results. For gel source data see Supplementary Fig. 1.

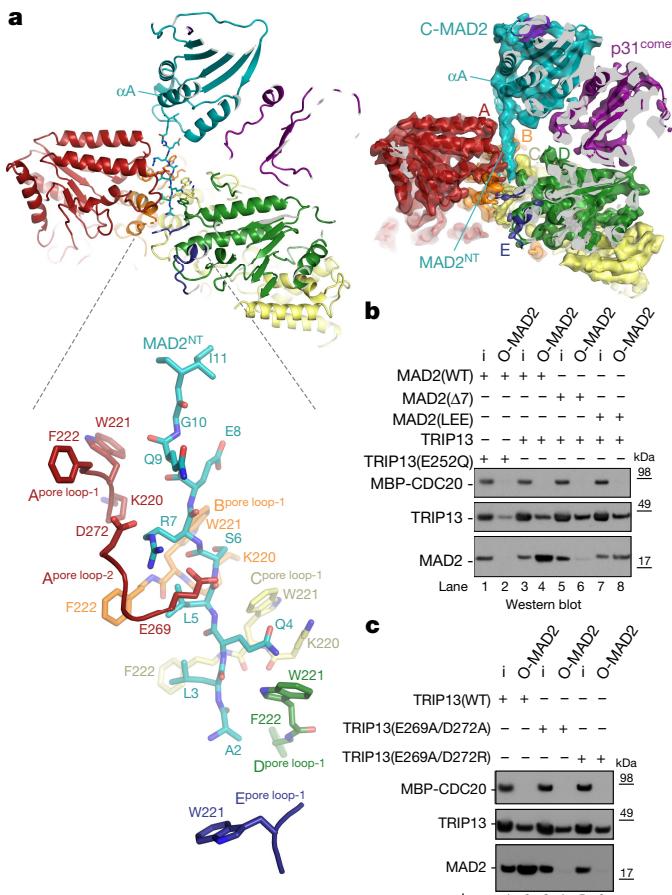


Fig. 3 | Interaction between C-MAD2^{NT} and the TRIP13 pore loops. **a**, Top, overview of the TRIP13-p31-substrate complex showing MAD2^{NT} inserted into the TRIP13 pore (left) with corresponding electron microscopy density map (right). Bottom, detailed representation of MAD2^{NT} residues interacting with TRIP13 pore loop residues. **b**, Western blot showing input (i) and O-MAD2 fractions of the disassembly of the CDC20-C-MAD2 complex by TRIP13(E253Q) (lanes 1, 2) and wild-type TRIP13 (lanes 3–8). MAD2 levels and loading controls for TRIP13 and CDC20 were detected using their respective antibodies. Mutants of MAD2^{NT} (lanes 5–8) affect the amount of O-MAD2 released from the CDC20-C-MAD2 complex. **c**, Mutation of D269 and E272 of pore loop-2 of TRIP13 ablates TRIP13-p31^{comet}-catalysed remodelling of MAD2 (Extended Data Fig. 2c). Experiments in **b**, **c** were performed in triplicate with similar results.

In cells, although wild-type MAD2 is in the open state^{18,20}, the $\Delta N5$ -MAD2 mutant exists 50% in the closed state and impairs SAC silencing¹⁸. Deletion of the first ten residues resulted in the absence of cellular O-MAD2, and caused a more severe mitotic defect¹⁸.

AAA+ ATPases are molecular motors that convert the chemical energy of ATP hydrolysis into mechanical energy²⁴. To explore how TRIP13 ATPase activity induces conformational changes in C-MAD2, we modelled the likely structure of the TRIP13-p31-substrate complex after the first cycle of catalysis (Fig. 4b, Extended Data Fig. 9, Supplementary Video 2). The largest structural change involves the seam monomer F and neighbouring monomer E. Monomer F adopts an active (ATP-bound) conformation, translates by about 12 Å to the top of the AAA+ spiral and establishes contacts with MAD2^{NT} two residues along relative to monomer A (Fig. 4a, b, Extended Data Fig. 9b). Thus, monomers E and F (E¹ and F¹ in Fig. 4b and Extended Data Fig. 9b; superscript '1' denotes post catalytic cycle 1) climb relative to C-MAD2, a process that pushes monomer F¹ onto the α A and α C helices of C-MAD2 (Fig. 4b, Extended Data Fig. 9b, c). The new

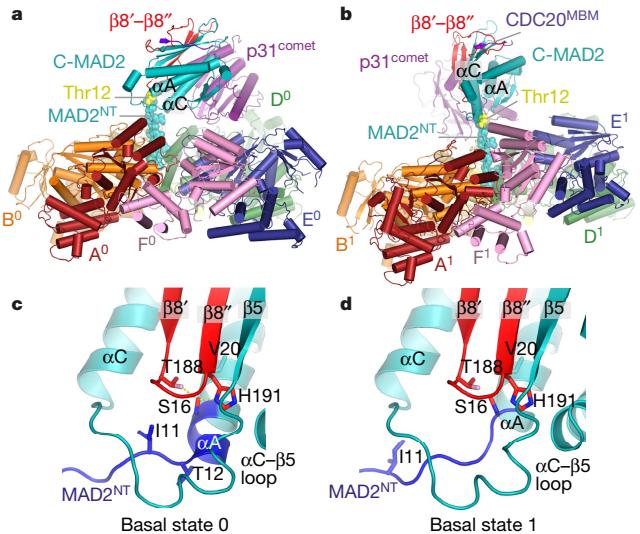


Fig. 4 | Sequential catalytic cycles of TRIP13 remodel MAD2. **a**, **b**, Overall views of the TRIP13-p31-substrate complex before catalysis (**a**; pre-catalytic: basal state 0) and after the first catalytic cycle (**b**; basal state 1). TRIP13 subunit superscripts denote the catalytic cycle. Thr12 of MAD2^{NT} is coloured yellow to indicate the boundary with α A in basal state 0. **c**, **d**, After the first catalytic cycle, one turn of the α A helix unwinds (compare **c** and **d**). This disrupts contacts between the N terminus of the α A helix and the β 8'- β 8'' hairpin. The N-terminal region of C-MAD2 that differs between O-MAD2 and C-MAD2 is in blue (Extended Data Fig. 6b, c). Further structural details are in Extended Data Fig. 9.

conformation of monomer E (E¹ in Fig. 4b), not only differs, thereby disrupting the TRIP13-p31^{comet} interface, but also severely clashes with p31^{comet} (Extended Data Fig. 9c, Supplementary Video 3). Owing to these steric clashes, the globular portion of p31-substrate moves vertically relative to MAD2^{NT} held within the TRIP13 pore by subunits A¹ to D¹ (Extended Data Fig. 9a, d). An anticlockwise rotation of the globular portion of p31-substrate allows p31^{comet} to engage the reset TRIP13 interface 60° anticlockwise. Were the entire p31-substrate (including MAD2^{NT}) to shift, the Leu13 C α atom (the start of α A of C-MAD2) would be displaced by 7.6 Å after the first catalytic step (Extended Data Fig. 9i). However, because MAD2^{NT} is gripped by the TRIP13 pore, fixing Leu13 C α , the α A helix unwinds to stretch the polypeptide chain (Extended Data Fig. 9f–j). Modelling indicates that each round of catalysis causes a single α -helical turn to unwind into an extended conformation, coupled with progression of MAD2^{NT} through the TRIP13 pore by two residues (Fig. 4, Extended Data Fig. 9g–j, Supplementary Video 2).

Unwinding the α A helix suggests that the conversion of C-MAD2 to O-MAD2 occurs through disruption of contacts with the adjacent β 8'- β 8'' hairpin, including hydrogen bonds linking Ser16 of α A with Thr188 and His191 of β 8'- β 8'' (Fig. 4c, d, Supplementary Video 2). As these hydrogen bonds stabilize C-MAD2^{18,25}, their loss would displace β 8'- β 8''. Opening the buckle (β 8'- β 8'' pairing with β 5) releases the safety belt structure that traps CDC20 to C-MAD2, liberating CDC20 (Extended Data Fig. 6d). Restructuring of β 8'- β 8'' and the C-MAD2 C terminus weakens interactions with p31^{comet} that bind C-MAD2^{19,21,22} (Extended Data Fig. 6e). As proposed¹⁷, this coordinates remodelling of C-MAD2 with release from TRIP13, preventing substantial unfolding of MAD2. Release of MAD2^{NT} would promote conversion to O-MAD2 because of the role of β 1 in stabilizing O-MAD2^{19,22,23,26}. The tendency of the TRIP13 hexamer to disassemble¹⁸, especially in the absence of p31-substrate, may provide a mechanism for release of MAD2^{NT}. Working on the estimate that 8–10 ATP molecules are hydrolysed for each C-MAD2 molecule converted¹⁷, the conformational transition of C-MAD2 would involve unwinding of no more than the α A helix.

p31-substrate stimulates TRIP13 ATPase activity¹⁷, indicating that it promotes an activating conformational change of TRIP13.

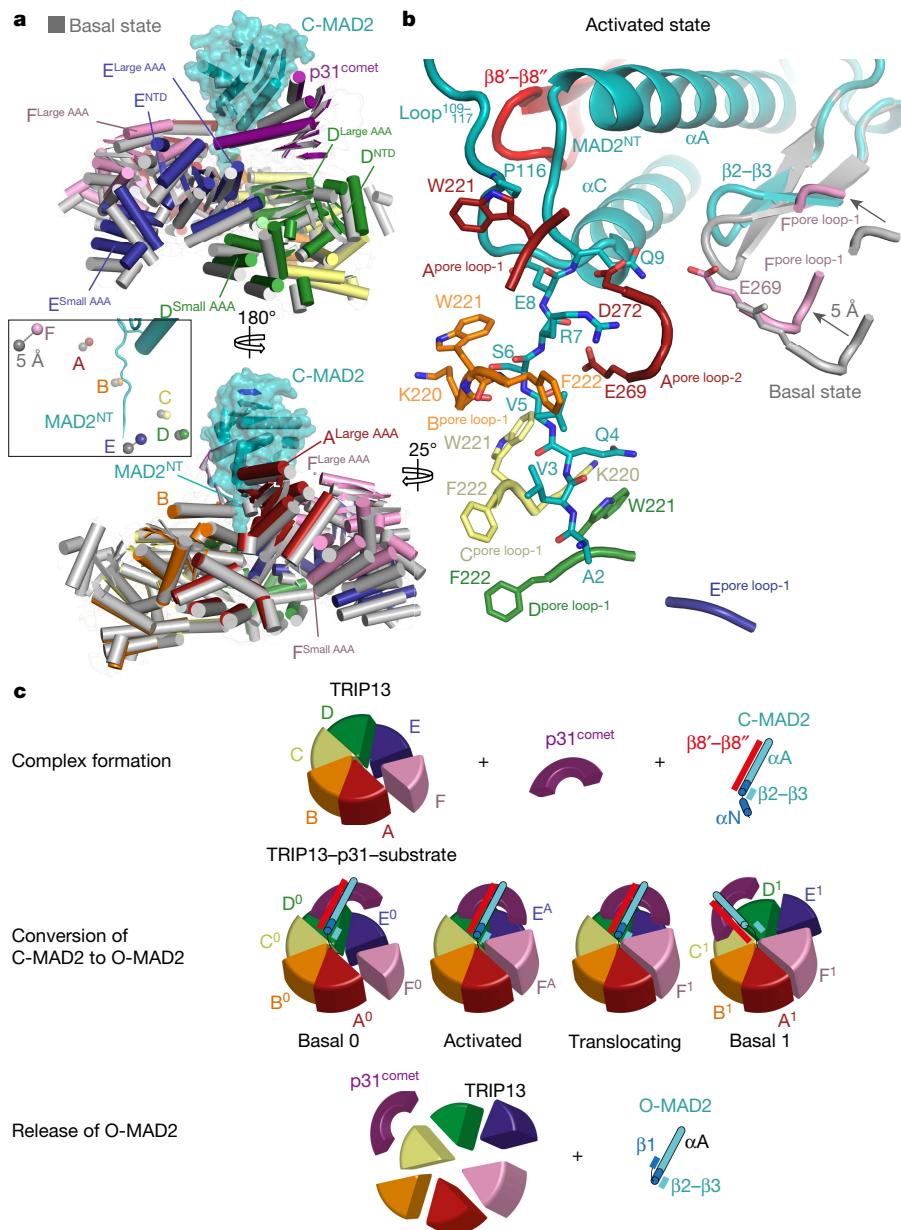


Fig. 5 | Differences between basal and activated states of TRIP13-p31-substrate. **a**, Two views comparing the basal state (grey) and activated state (coloured as in Fig. 1) of TRIP13-p31-substrate (superimposition on C-MAD2). Left inset shows displacement of reference pore loop residues between basal and active states. The displacement grows from the p31^{comet}-contacting monomer D to the F monomer, with F featuring the largest displacement (5 Å). **b**, The conformational changes of

monomer F pore loops and the C-MAD2 β2-β3 hairpin in the active state are shown relative to their positions in the basal state (grey). **c**, Schematic of the proposed mechanism of TRIP13-p31^{comet}-catalysed remodelling of O-MAD2 (Supplementary Video 2). The remodelled structural elements of MAD2 are indicated. E^A and F^A, activated conformations of monomers E and F, respectively. The O-MAD2 products dissociate from p31^{comet}.

The conformational differences in the TRIP13-p31-substrate complex between classes 2 and 3 suggest that class 3 is an activated state primed for ATP hydrolysis. Compared with class 2, TRIP13 monomers D-F have changed their orientations relative to monomers A-C, with the conformational change progressively increasing from D to F and resulting in their shift towards p31-substrate (Fig. 5a, Supplementary Video 4). These changes in TRIP13 are accompanied by a small rotation of p31^{comet}-C-MAD2 so that p31^{comet} itself moves closer to TRIP13 monomers D and E. Monomer F shows the largest displacement towards the p31-substrate module, allowing its pore loop-2 to approach MAD2^{NT} (Fig. 5a,b). The effect of the p31-substrate-induced conformational change in TRIP13 is to move monomers E and F along the axis of translocation.

On conversion of O-MAD2 to C-MAD2 the β2-β3 hairpin rearranges. Mutations of the β2-β3 hairpin suppress this conversion²⁷.

In the activated state, the approaching F monomer shifts the β2-β3 hairpin towards αA, perturbing its interactions with the C-MAD2 core (Fig. 5b). Thus, it is possible that the energy of p31-substrate binding to TRIP13 is converted into torsional energy within the TRIP13 ring that destabilizes C-MAD2 core interactions with the β2-β3 hairpin.

The proposed TRIP13-catalysed rotary motion of the globular region of p31^{comet}-C-MAD2 relative to the fixed MAD2^{NT} presents a mechanism to explain how local unwinding of a secondary structural element remodels tertiary and quaternary structures (Fig. 5c).

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0281-1>.

Received: 14 November 2017; Accepted: 16 May 2018;
Published online: 04 July 2018

1. Musacchio, A. The molecular biology of spindle assembly checkpoint signaling dynamics. *Curr. Biol.* **25**, R1002–R1018 (2015).
2. Alfieri, C., Zhang, S. & Barford, D. Visualizing the complex functions and mechanisms of the anaphase promoting complex/cyclosome (APC/C). *Open Biol.* (2017).
3. Kulukian, A., Han, J. S. & Cleveland, D. W. Unattached kinetochores catalyze production of an anaphase inhibitor that requires a Mad2 template to prime Cdc20 for BubR1 binding. *Dev. Cell* **16**, 105–117 (2009).
4. Faesen, A. C. et al. Basis of catalytic assembly of the mitotic checkpoint complex. *Nature* **542**, 498–502 (2017).
5. Ji, Z., Gao, H., Jia, L., Li, B. & Yu, H. A sequential multi-target Mps1 phosphorylation cascade promotes spindle checkpoint signaling. *eLife* **6**, e22513 (2017).
6. Habu, T., Kim, S. H., Weinstein, J. & Matsumoto, T. Identification of a MAD2-binding protein, CMT2, and its role in mitosis. *EMBO J.* **21**, 6419–6428 (2002).
7. Eytan, E. et al. Disassembly of mitotic checkpoint complexes by the joint action of the AAA-ATPase TRIP13 and p31^{comet}. *Proc. Natl Acad. Sci. USA* **111**, 12019–12024 (2014).
8. Wang, K. et al. Thyroid hormone receptor interacting protein 13 (TRIP13) AAA-ATPase is a novel mitotic checkpoint-silencing protein. *J. Biol. Chem.* **289**, 23928–23937 (2014).
9. Teichner, A. et al. p31^{comet} promotes disassembly of the mitotic checkpoint complex in an ATP-dependent process. *Proc. Natl Acad. Sci. USA* **108**, 3187–3192 (2011).
10. Westhorpe, F. G., Tighe, A., Lara-Gonzalez, P. & Taylor, S. S. p31^{comet}-mediated extraction of Mad2 from the MCC promotes efficient mitotic exit. *J. Cell Sci.* **124**, 3905–3916 (2011).
11. Reddy, S. K., Rape, M., Margansky, W. A. & Kirschner, M. W. Ubiquitination by the anaphase-promoting complex drives spindle checkpoint inactivation. *Nature* **446**, 921–925 (2007).
12. Foster, S. A. & Morgan, D. O. The APC/C subunit Mnd2/Apc15 promotes Cdc20 autoubiquitination and spindle assembly checkpoint inactivation. *Mol. Cell* **47**, 921–932 (2012).
13. Mansfeld, J., Collin, P., Collins, M. O., Choudhary, J. S. & Pines, J. APC15 drives the turnover of MCC-CDC20 to make the spindle assembly checkpoint responsive to kinetochore attachment. *Nat. Cell Biol.* **13**, 1234–1243 (2011).
14. Uzunova, K. et al. APC15 mediates CDC20 autoubiquitylation by APC/C(MCC) and disassembly of the mitotic checkpoint complex. *Nat. Struct. Mol. Biol.* **19**, 1116–1123 (2012).
15. Jia, L. et al. Defining pathways of spindle checkpoint silencing: functional redundancy between Cdc20 ubiquitination and p31^{comet}. *Mol. Biol. Cell* **22**, 4227–4235 (2011).
16. Eytan, E., Sityr-Shevah, D., Teichner, A. & Hershko, A. Roles of different pools of the mitotic checkpoint complex and the mechanisms of their disassembly. *Proc. Natl Acad. Sci. USA* **110**, 10568–10573 (2013).
17. Ye, Q. et al. TRIP13 is a protein-remodeling AAA+ ATPase that catalyzes MAD2 conformation switching. *eLife* **4**, e07367 (2015).
18. Ye, Q. et al. The AAA+ ATPase TRIP13 remodels HORMA domains through N-terminal engagement and unfolding. *EMBO J.* **36**, 2419–2434 (2017).
19. Yang, M. et al. p31^{comet} blocks Mad2 activation through structural mimicry. *Cell* **131**, 744–755 (2007).
20. Ma, H. T. & Poon, R. Y. C. TRIP13 regulates both the activation and inactivation of the spindle-assembly checkpoint. *Cell Reports* **14**, 1086–1099 (2016).
21. Xia, G. et al. Conformation-specific binding of p31^{comet} antagonizes the function of Mad2 in the spindle checkpoint. *EMBO J.* **23**, 3133–3143 (2004).
22. Mapelli, M. et al. Determinants of conformational dimerization of Mad2 and its inhibition by p31^{comet}. *EMBO J.* **25**, 1273–1284 (2006).
23. Mapelli, M., Massimiliano, L., Santaguida, S. & Musacchio, A. The Mad2 conformational dimer: structure and implications for the spindle assembly checkpoint. *Cell* **131**, 730–743 (2007).
24. Lyubimov, A. Y., Strycharski, M. & Berger, J. M. The nuts and bolts of ring-translocase structure and mechanism. *Curr. Opin. Struct. Biol.* **21**, 240–248 (2011).
25. Yang, M. et al. Insights into mad2 regulation in the spindle checkpoint revealed by the crystal structure of the symmetric mad2 dimer. *PLoS Biol.* **6**, e50 (2008).
26. Luo, X. et al. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nat. Struct. Mol. Biol.* **11**, 338–345 (2004).
27. Hara, M., Özkan, E., Sun, H., Yu, H. & Luo, X. Structure of an intermediate conformer of the spindle checkpoint protein Mad2. *Proc. Natl Acad. Sci. USA* **112**, 11252–11257 (2015).

Acknowledgements This work was funded by MRC and CR-UK grants to D.B. C.A. is an EMBO Advanced Fellow. We thank A. Boland for comments on the manuscript; S. Chen, C. Savva and G. McMullan for help with EM data collection; and J. Grimmett and T. Darling for computing.

Reviewer information *Nature* thanks K. Corbett, H. Yu and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.A. cloned bacterially expressed p31^{comet}, MAD2 and TRIP13 wild type and mutant constructs, purified proteins, performed the protein complex reconstitutions and biochemical analysis and mutagenesis. C.A. prepared EM grids, analysed EM data and determined the three dimensional reconstructions. C.A. collected EM data with the help of L.C. C.A. fitted coordinates and built models. D.B. directed the project. C.A. and D.B. wrote the manuscript with input from L.C.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0281-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0281-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cloning, expression and purification of TRIP13 complexes. The codon-optimized DNA coding sequences (CDSs) of full-length human *TRIP13* (Q15645) and *p31^{comet}* (also known as *MAD2L1BP*; Q15013) were cloned individually into the bacterial pETM11 expression vector with an N-terminal TEV protease-cleavable His₆ tag. *MAD2* (also known as *MAD2L1*; Q13257) CDS was cloned into a custom-modified version of the pETM11 vector where the N-terminal His₆ tag was replaced by a C-terminal 3C-cleavable His₆ tag. TRIP13 and *p31^{comet}* mutants were generated using QuickChange and MAD2 mutants were generated using Gibson cloning. These include TRIP13: E269A plus D272A and E269R plus D272R, E253Q; MAD2: L56R7 to L56E7; and the N-terminal 7 and 9 residue deletions; *p31^{comet}*: deletion of R99, K100, K110, K111, K112, R122, K123 and K270A (TRIP13-binding site on monomer E). Proteins were individually expressed in the BL21 Star (DE3) *Escherichia coli* strain at 18 °C for 12 h. TRIP13, *p31^{comet}* and MAD2 were individually purified by Ni-NTA based protein purification (Ni-NTA agarose resin) followed by either TEV or 3C protease cleavage and gel filtration (Superdex 75 10/300 column). TRIP13 and *p31^{comet}* were further purified on a 6-ml Resource Q column (GE Healthcare Life Sciences) before size exclusion chromatography. Cloning, expression and purification of MCC and the APC/C-MCC complex were as described²⁸.

C-MAD2 complex remodelling experiments. Purified *p31^{comet}* and either wild-type or mutant MAD2 were incubated with purified maltose-binding protein (MBP)-CDC20 (Q13257)²⁸ for 2 h at 23 °C. The resulting *p31^{comet}*-C-MAD2-MBP-CDC20 complex was purified by gel filtration using a Superdex 200 10/300 column. For disassembly of either the *p31^{comet}*-C-MAD2-MBP-CDC20 complex or the MCC (both the MCC alone and MCC in the APC/C-MCC complex), 10–15 µmol of complex was incubated with a threefold molar excess of TRIP13 (either wild-type or the Q253E mutant, accounting for hexamerization of TRIP13), in a disassembly buffer (50 mM HEPES pH 8.0, 10 mM MgCl₂ and 3 mM ATP) in a total volume of 50 µl. In the reactions containing MCC and APC/C-MCC, a tenfold excess of *p31^{comet}* was also added. Following incubation at 23 °C for 10 min the disassembly reaction was injected into a ÄKTAmicro (GE Healthcare) with a running buffer of 20 mM HEPES pH 8.0, 300 mM NaCl, 10 mM MgCl₂ and 0.5 mM TCEP. O-MAD2 eluted in a later fraction than the complexes of C-MAD2 either with *p31^{comet}* or with both *p31^{comet}* and MBP-CDC20 (Extended Data Fig. 2a). O-MAD2 fractions were loaded (with input material) onto an SDS-PAGE gel and detected with a western blot using anti-MAD2 antibody (ab10691; Abcam). TRIP13, CDC20 and *p31^{comet}* were detected with antibodies against TRIP13 (sc-514285; Santa Cruz Biotechnology), CDC20 (sc-8358; Santa Cruz Biotechnology) and *p31^{comet}* (MABE451; Merck), respectively.

To test dissociation of MCC into C-MAD2-CDC20 and BUBR1-BUB3 (BUBR1: O60566; BUB3: O43684) sub-complexes by (i) the combined action of TRIP13(E253Q) and *p31^{comet}*, (ii) TRIP13(E253Q) alone and (iii) *p31^{comet}* alone, 30 µmol MCC was mixed with (i) a 1.1-fold excess of TRIP13(E253Q) hexamer and a tenfold excess of *p31^{comet}*, (ii) a 1.1-fold excess of TRIP13(E253Q) or (iii) a tenfold excess of *p31^{comet}* in buffer: 50 mM HEPES pH 8.0, 10 mM MgCl₂ and 3 mM ATP. The mixture was injected into a Superdex 200 10/300 gel filtration column fitted to an ÄKTAmicro (GE Healthcare) with a running buffer of 20 mM HEPES pH 8.0, 300 mM NaCl, 10 mM MgCl₂ and 0.5 mM TCEP and ATP. Peak eluted fractions were analysed using Coomassie-stained gels.

Reconstitution of TRIP13-p31-substrate complex for EM analysis. A ternary complex composed of *p31^{comet}*-C-MAD2-CDC20 was purified by Ni-NTA column chromatography starting with a co-lysate of Hi-5 insect cells²⁸ co-expressing C-MAD2-MBP-CDC20 with BL21 Star (DE3) cells expressing *p31^{comet}*, followed by TEV cleavage and gel filtration (Superdex 200 10/300 column). Purified TRIP13(Q253E) was incubated with equimolar amounts of the *p31^{comet}*-C-MAD2-CDC20 complex (accounting for hexamerization of TRIP13) in an assembly reaction buffer (50 mM HEPES pH 8.0, 10 mM MgCl₂ and 5 mM ATPγS) for 45 min at 23 °C. The 50-µl assembly reaction was injected into a Superdex 200 5/150 GL column fitted to an ÄKTAmicro (GE Healthcare) with a running buffer containing 20 mM HEPES pH 8.0, 300 mM NaCl, 10 mM MgCl₂, 0.5 mM TCEP and 0.3 mM ATPγS. The complex eluted in a peak fraction at a concentration of 0.3 mg/ml and was used to prepare cryo-EM grids.

Electron microscopy. We applied 2.5 µl of the TRIP13-p31-substrate complex eluted fraction to Quantifoil Holey carbon R1.2/1.3 Au 300 grids, treated with a 9:1 argon:oxygen plasma for 30 s before use. The grids were incubated for 30 s at 4 °C and 100% humidity and then blotted for 6 s and plunged into liquid ethane using an FEI Vitrobot III. Specimens were imaged using both EPU software (FEI) and Serial EM (Mastronarde Group) at a nominal magnification of 81,000 \times , yielding a pixel size of 1.43 Å at specimen level on an FEI Titan Krios electron microscope operating at 300 kV accelerating voltage. Zero-energy-loss micrographs were recorded using a Gatan K2-Summit direct electron detector executed in counting mode at the end of a Gatan GIF-Quantum energy filter with a slit width of 20 eV. Images were collected at a dose rate of ~2.6 electrons per Å² per s. Exposures of 16 s were

dose-fractionated into 20 movie frames with a total dose of ~40 electrons per Å². Defocus values in the final data set ranged from −2.0 to −3.6 µm.

Image processing. Movie frames were aligned using MotionCor2²⁹ before subsequent processing. Contrast transfer function parameters were calculated using Gctf³⁰. Particles in 140 × 140 pixels were selected automatically using Gautomatch (K. Zhang) with an inter-particle distance cutoff of 150 Å. The initial 2D references for the automated picking were created by manual picking. Picked particles were extracted and processed in RELION 2.1³¹. After three rounds of 2D classification in which 2D classes with poor structural features were removed, 30–40% of particle classes were finally selected for 3D refinement. The initial model for 3D refinement was generated ab initio using the random start-up procedure in IMAGIC-4D³². Overall 3D refinement yielded a 3D reconstruction with 4.5 Å resolution. To further improve the EM density map of TRIP13, a soft mask including the most rigid TRIP13 monomers A/B/C/D (TRIP13^{A/B/C/D}) was used for local alignment during refinement. This improved the resolution to 4.3 Å for TRIP13^{A/B/C/D}. All resolution estimations were from gold-standard FSC calculations to avoid over-fitting and reported resolutions are based on the FSC = 0.143 criterion³³. Final FSC curves and sharpening of density maps was performed as described³¹.

3D classification with a global search and a sampling angular interval of 7.5° allowed the identification of apo (class 1: 20% of total) and *p31*-substrate-bound (80% of total) TRIP13 complexes. The TRIP13-p31-substrate class was further classified yielding the TRIP13-p31-substrate basal (class 2: 22%) and TRIP13-p31-substrate activated states (class 3: 24%) and a reconstruction (34%) with resemblance to the TRIP13-p31-substrate basal state but poorer alignment accuracy.

Map visualization. Figures and videos were generated using PyMOL (Molecular Graphics System, 2.0.3, Schrödinger, LLC) and Chimera³⁴.

Model building. Initial fitting and superposition of human TRIP13¹⁸ and *p31^{comet}*-C-MAD2-MBP¹⁹ were performed using Chimera³⁴. Model building and flexible fitting were performed in COOT³⁵. TRIP13 pore loops-1 and -2 model building was performed ab initio using the crystal structure of the PAN regulatory particle³⁶ as an initial reference using our highest resolution maps. TRIP13 loop 111–121 was built as an idealized polyAla chain fitting to the corresponding density. The *p31^{comet}* α3–4 loop was built as polyAla chain fitting the corresponding density of a TRIP13-p31-substrate map obtained after focused classification of the N-terminal domain of TRIP13 monomer E with additional guidance from published cross-linking data restraints¹⁸. MAD2^{NT} was built by flexible fitting of MAD2²⁵. Model refinement of the TRIP13-p31-substrate structure was performed with PHENIX³⁷ using 3D classes 2 and 3.

Modelling TRIP13-p31^{comet}-C-MAD2-CDC20 complexes after catalytic cycles. See Extended Data Fig. 9 and Supplementary Videos 2 and 3. To model TRIP13 after the first catalytic cycle (that is, basal state 1), we superimposed monomer B onto monomer A (thus new A¹ was formerly B⁰, new F¹ was formerly A⁰ and new E¹ was F⁰ and so on, where superscript '0' denotes the pre-catalytic state and superscript '1' denotes the first post-catalytic state). C-MAD2^{NT} (here defined as residues 2–12) was modelled based on its conformation in the pre-catalytic conformation (that is, unchanged). The globular portion of *p31^{comet}*-C-MAD2 (all residues of *p31^{comet}* and residues 13–205 of C-MAD2) were re-positioned onto the new C'-D¹ interface (in effect maintaining the same contacts with the shifted D⁰-E⁰ subunits) to avoid clashes with the repositioned E¹ and F¹. Because C-MAD2^{NT} remained fixed, but the globular portion of *p31^{comet}*-C-MAD2 shifted by 7.6 Å (owing to the vertical movement and 60° rotation (Extended Data Fig. 9j)), the Cα atoms of Thr12 of C-MAD2^{NT} and Leu13 in the globular portion of C-MAD2 were separated by 11 Å (Extended Data Fig. 9h). Thr12 and Leu13 were reconnected by unwinding the first turn of the C-MAD2 αA helix (residues 12–17) with the coordinates refined in COOT³⁵. The same procedure was applied to remodel the TRIP13-p31^{comet}-C-MAD2-CDC20 complex basal state 2 (post catalytic cycle 2) (Supplementary Video 2). The calculation that one helical turn is unwound per catalytic cycle is based on the following reasoning:

The axial distance between adjacent residues in an α-helix is 1.5 Å.

The axial distance between adjacent residues in a β-sheet is 3.5 Å.

Thus the difference in axial distance between adjacent residues in an α-helix and a β-sheet is 2 Å per residue. Stretching an α-helix by 7.6 Å by forming an extended chain requires 7.6 Å/2 Å per residue = 3.8 residues = 1.06 helical turn (assuming 3.6 residues per turn).

Structural conservation was calculated using CONSURF³⁸. The interface surface area between TRIP13 and *p31^{comet}*-C-MAD2 including MAD2^{NT} was calculated using the PISA protein interface server³⁹. Protein sequence alignment was performed using Jalview⁴⁰.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Figure source data are provided in Supplementary Fig. 1. Cryo-EM maps of the TRIP13-p31^{comet}-substrate complex have been deposited with the Electron Microscopy Data Bank under accession number EMD-4166. Atomic coordinates of the TRIP13-p31^{comet}-substrate complex have been deposited with the RCSB Protein Databank under entry ID 6F0X.

28. Alfieri, C. et al. Molecular basis of APC/C regulation by the spindle assembly checkpoint. *Nature* **536**, 431–436 (2016).

29. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).

30. Zhang, K. Gctf: Real-time CTF determination and correction. *J. S. Biol.* **193**, 1–12 (2016).

31. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D Struct. Biol.* **73**, 496–502 (2017).

32. Afanasyev, P. et al. Single-particle cryo-EM using alignment by classification (ABC): the structure of *Lumbricus terrestris* haemoglobin. *IUCrJ* **4**, 678–694 (2017).

33. Chen, S. et al. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24–35 (2013).

34. Yang, Z. et al. UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J. Struct. Biol.* **179**, 269–278 (2012).

35. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

36. Zhang, F. et al. Structural insights into the regulatory particle of the proteasome from *Methanocaldococcus jannaschii*. *Mol. Cell* **34**, 473–484 (2009).

37. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

38. Landau, M. et al. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33**, W299–W302 (2005).

39. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

40. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

41. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

42. Chao, W. C., Kulkarni, K., Zhang, Z., Kong, E. H. & Barford, D. Structure of the mitotic checkpoint complex. *Nature* **484**, 208–213 (2012).

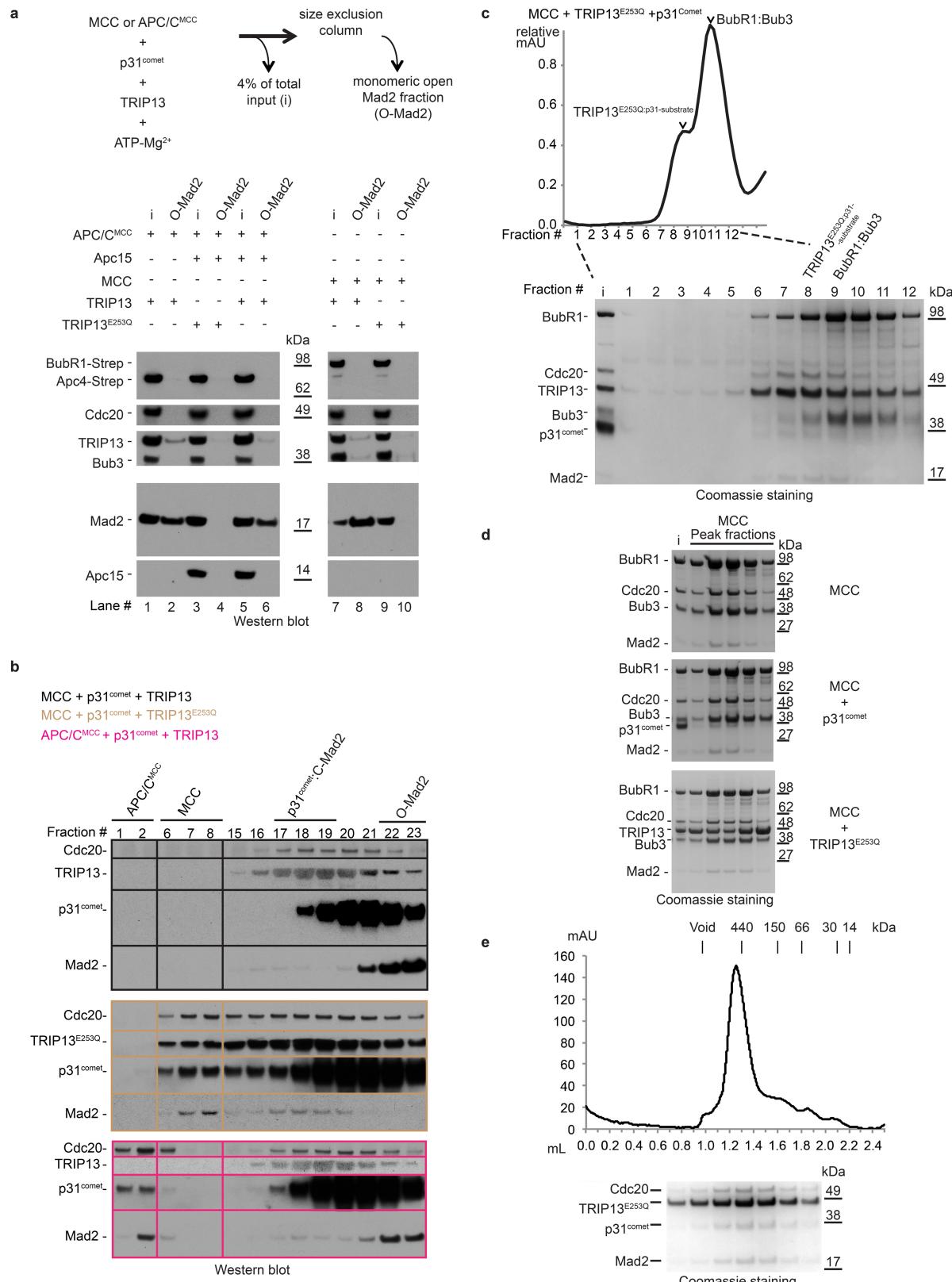
43. Luo, X., Tang, Z., Rizzo, J. & Yu, H. The Mad2 spindle checkpoint protein undergoes similar major conformational changes upon binding to either Mad1 or Cdc20. *Mol. Cell* **9**, 59–71 (2002).

44. Sironi, L. et al. Crystal structure of the tetrameric Mad1-Mad2 core complex: implications of a ‘safety belt’ binding mechanism for the spindle checkpoint. *EMBO J.* **21**, 2496–2506 (2002).

45. Ripstein, Z. A., Huang, R., Augustyniak, R., Kay, L. E. & Rubenstein, J. L. Structure of a AAA+ unfoldase in the process of unfolding substrate. *eLife* **6**, <https://doi.org/10.7554/eLife.25754> (2017).

46. Monroe, N., Han, H., Shen, P. S., Sundquist, W. I. & Hill, C. P. Structural basis of protein translocation by the Vps4–Vta1 AAA ATPase. *eLife* **6**, <https://doi.org/10.7554/eLife.24487> (2017).

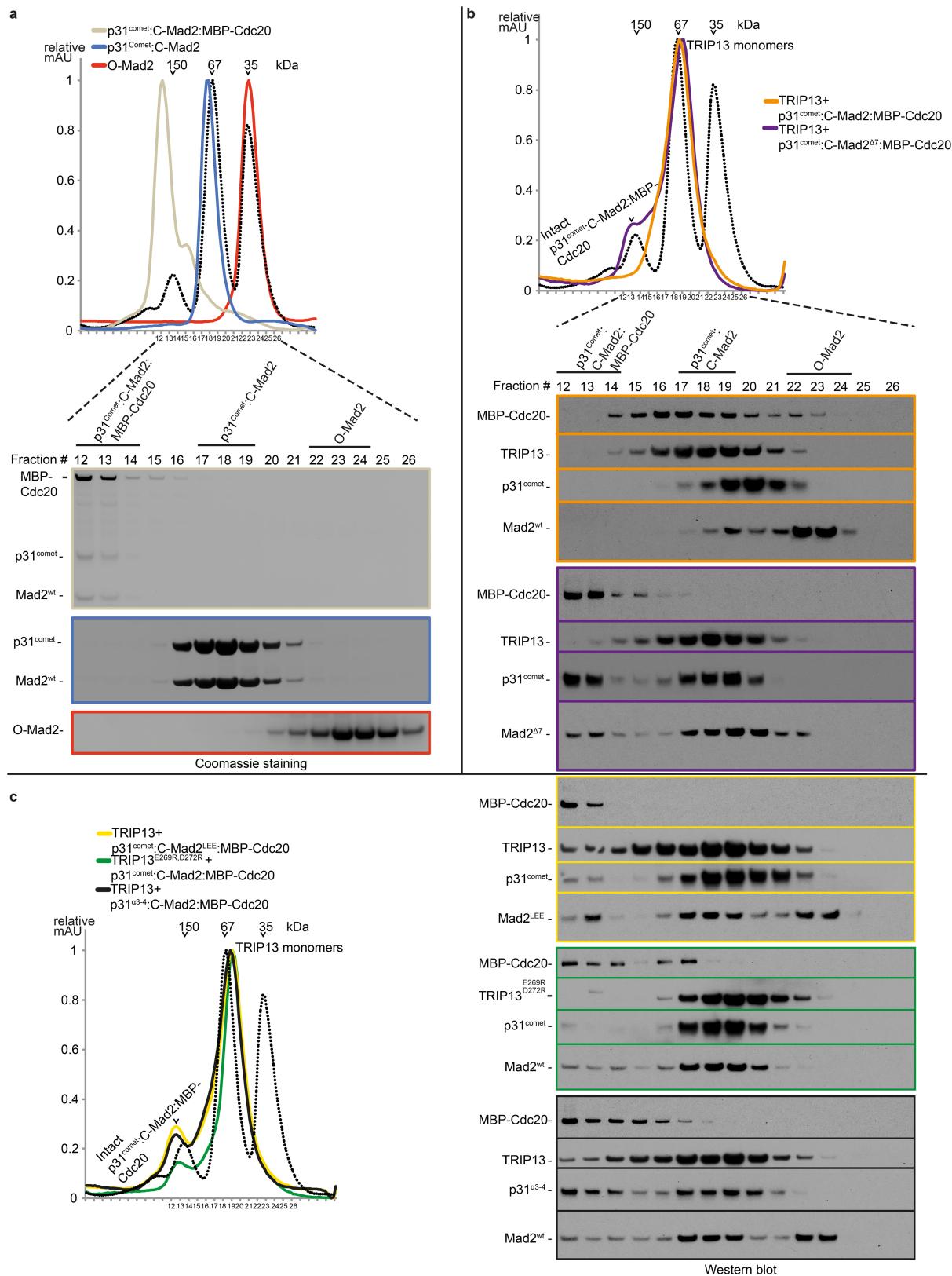
47. Gates, S. N. et al. Ratchet-like polypeptide translocation mechanism of the AAA+ disaggregase Hsp104. *Science* **357**, 273–279 (2017).



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Biochemical characterization of MAD2-containing complexes with TRIP13. **a**, TRIP13 and p31^{comet} can extract O-MAD2 from MCC and APC/C-MCC regardless of APC15. Western blot showing the disassembly reactions together with the respective input material (i) of APC/C-bound MCC (APC/C-MCC), in either the absence (lanes 1–2) or presence of APC15 (lanes 3–6) and MCC alone (lanes 7–10) (experimental design is shown on top). Negative control reactions (lanes 3, 4 and 9, 10) were performed with the TRIP13(E253Q) mutant. MAD2 levels were detected with an anti-MAD2 antibody. Loading controls of APC4-STREP, APC15, CDC20, TRIP13 and BUB3 were detected with antibodies specific for STREP, APC15, CDC20, TRIP13 and BUB3, respectively. BUBR1 (lanes 7–10) was detected with an anti-STREP antibody. **b**, Western blots showing eluted size exclusion (Superdex 200 10/300 column) fractions of the MCC and MAD2 remodelling reactions catalysed by TRIP13–p31^{comet} in the context of free MCC and APC/C-MCC. The fractions corresponding to (i) APC/C-MCC, (ii) MCC,

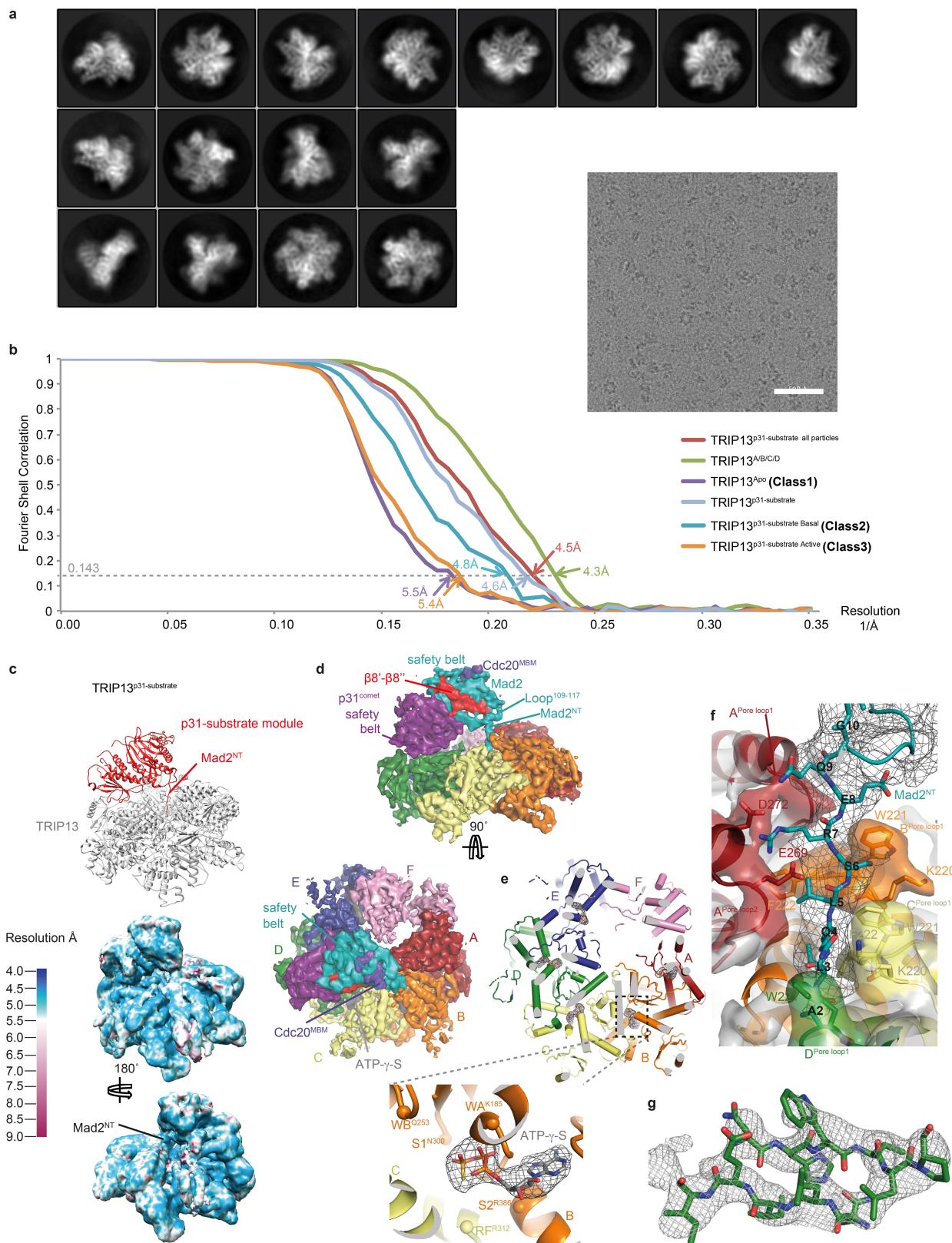
(iii) p31^{comet}–C-MAD2 and (iv) monomeric C-MAD2 are shown. A reference gel for size exclusion column elution fractions corresponding to p31^{comet}–C-MAD2–CDC20–MBP, p31^{comet}–C-MAD2 and monomeric C-MAD2 is shown in Extended Data Fig. 2a. **c**, Analysis of TRIP13–p31^{comet} complexes with the MCC using size exclusion chromatography in the presence of ATP. Coomassie-stained gel showing the gel filtration fractions (chromatogram above) of p31^{comet}–TRIP13 complexes in complex with MCC (fraction 8 is the p31^{comet}–TRIP13 complex with C-MAD2–CDC20 and fraction 9 is the BUBR1–BUB3 complex). Input material (i) is shown on the left. **d**, MCC binds p31^{comet} and not TRIP13 alone. Coomassie-stained gel showing the gel filtration fractions of the MCC (top gel) and in the presence of p31^{comet} (middle gel) and TRIP13(E253Q) (lower gel). **e**, Chromatogram (top) and SDS-PAGE (bottom) of the gel filtration performed with the TRIP13(E253Q)–p31^{comet}–C-MAD2–CDC20 complex in the presence of ATP γ S. Experiments in **a–e** were performed in triplicate with similar results. See Supplementary Fig. 1 for gel source data.



Extended Data Fig. 2 | See next page for caption.

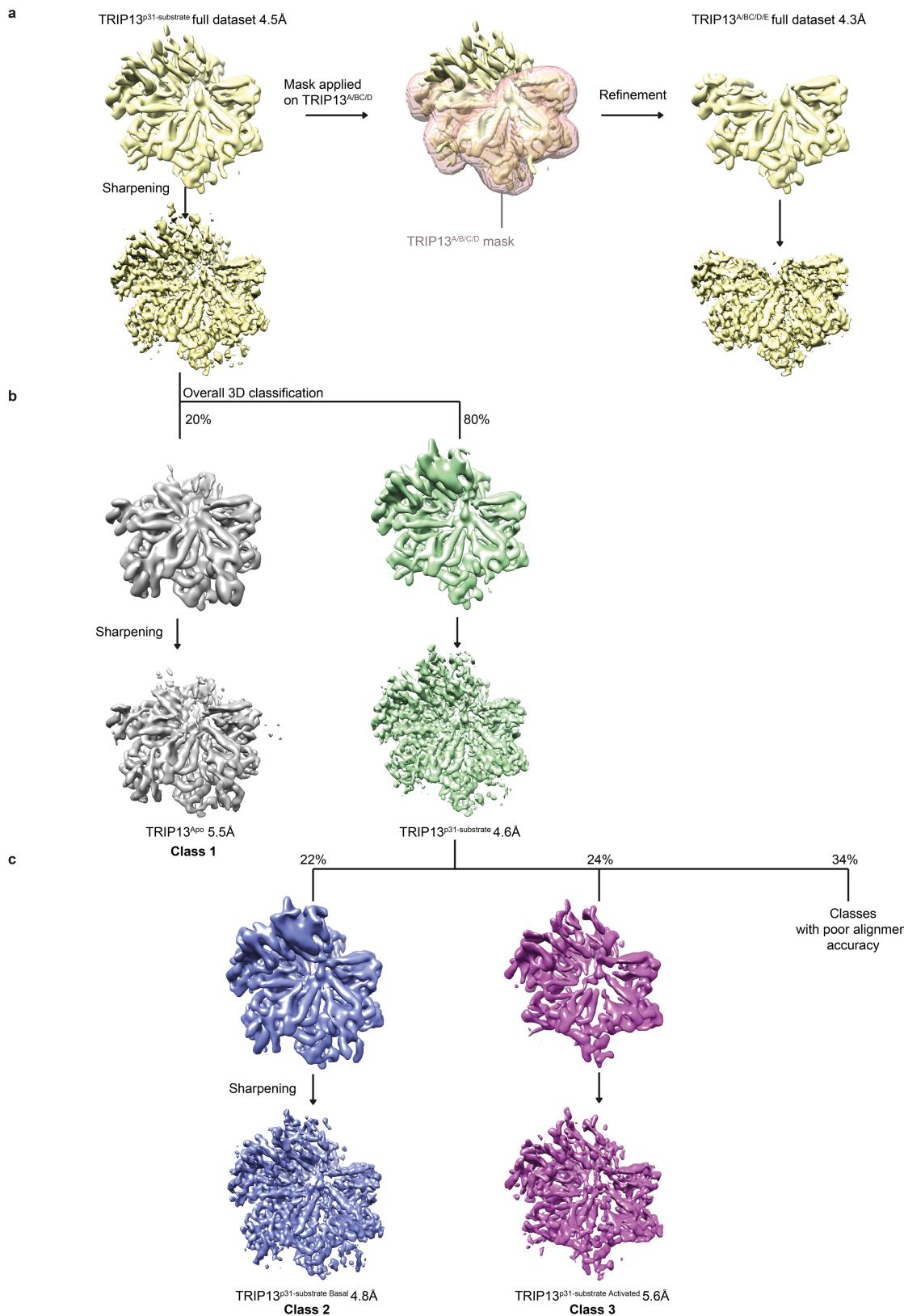
Extended Data Fig. 2 | Biochemical assay for TRIP13–p31^{comet}–catalysed O-MAD2 generation. Shown are size exclusion (Superdex 200 10/300 column) chromatograms and corresponding Coomassie-stained gels for the Mad2 remodelling reaction catalysed by TRIP13–p31^{comet}. **a**, Reference chromatograms and Coomassie-stained gels for (i) p31^{comet}–C-MAD2–MBP–CDC20 (brown trace), (ii) p31^{comet}–C-MAD2 (blue trace) and (iii) monomeric O-MAD2 (red trace). Chromatograms and gels are colour-coded. Monomeric O-MAD2 elutes in fractions 22–24, whereas p31^{comet}–C-MAD2 elutes in fractions 17–19. **b**, Western blots

for the products of the reaction of TRIP13 with (i) wild-type p31^{comet} and MAD2 (orange trace) and (ii) wild-type p31^{comet} and mutant C-MAD2 (C-MAD2^{Δ7} (seven N-terminal residues deleted)). **c**, Western blots for the products of the reaction with (i) mutant C-MAD2^{LEE}, wild-type TRIP13 and p31^{comet} (yellow trace), (ii) mutant TRIP13(E269R/D272R) and wild-type p31^{comet} and C-MAD2 (green trace) and (iii) mutant p31^{comet} α3–4 and wild-type TRIP13 and C-MAD2 (black trace). Experiments in **a–c** were performed in triplicate with similar results. See Supplementary Fig. 1 for gel source data.



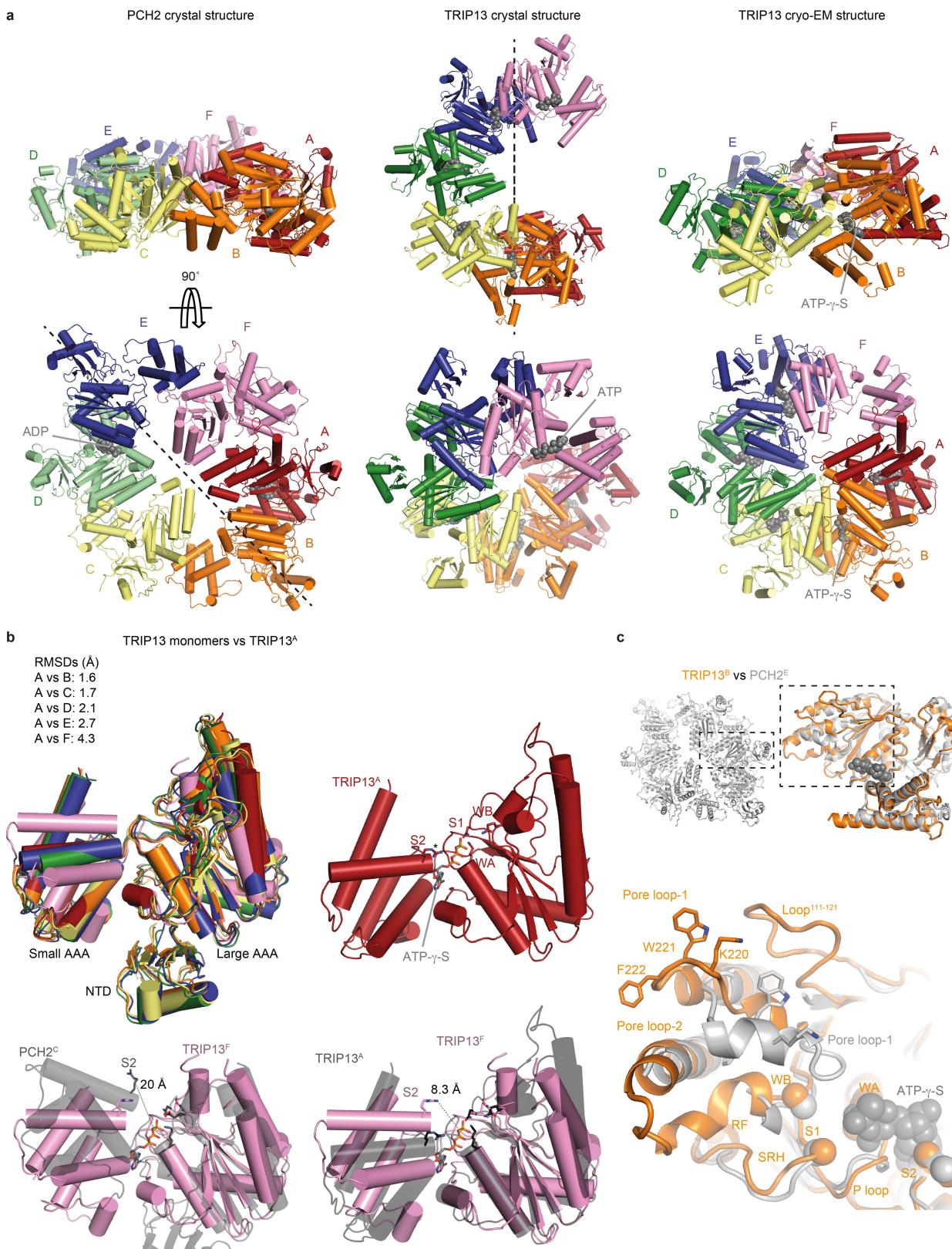
Extended Data Fig. 3 | Cryo-EM analysis and resolution of TRIP13 complexes in this study. **a**, Left, gallery of 2D class averages of TRIP13-p31-substrate showing different views representative of 50 2D classes. Right, a typical cryo-EM micrograph of TRIP13-p31-substrate representative of 3,630 micrographs. **b**, Fourier shell correlation (FSC) curves are shown for all the cryo-EM reconstructions determined in this study. **c**, Local resolution maps calculated with RESMAP⁴¹ of the TRIP13-p31-substrate complex. **d**, Cryo-EM density of the TRIP13-p31-substrate

reconstruction shown as in Fig. 1b. **e, g**, Representative density quality for the ATP-γ-S (**e**) and β-strand (**g**). In **e**, critical residues for the TRIP13 catalytic site are indicated: WA (Walker A), WB (Walker B), S1 (sensor 1), S2 (sensor 2) and RF (Arg finger). **f**, Close up of the TRIP13 pore loops interacting with C-MAD2^{NT} from the cryo-EM density of TRIP13-p31-substrate. EM density for C-MAD2 shown in black mesh, TRIP13 in transparent coloured surface.



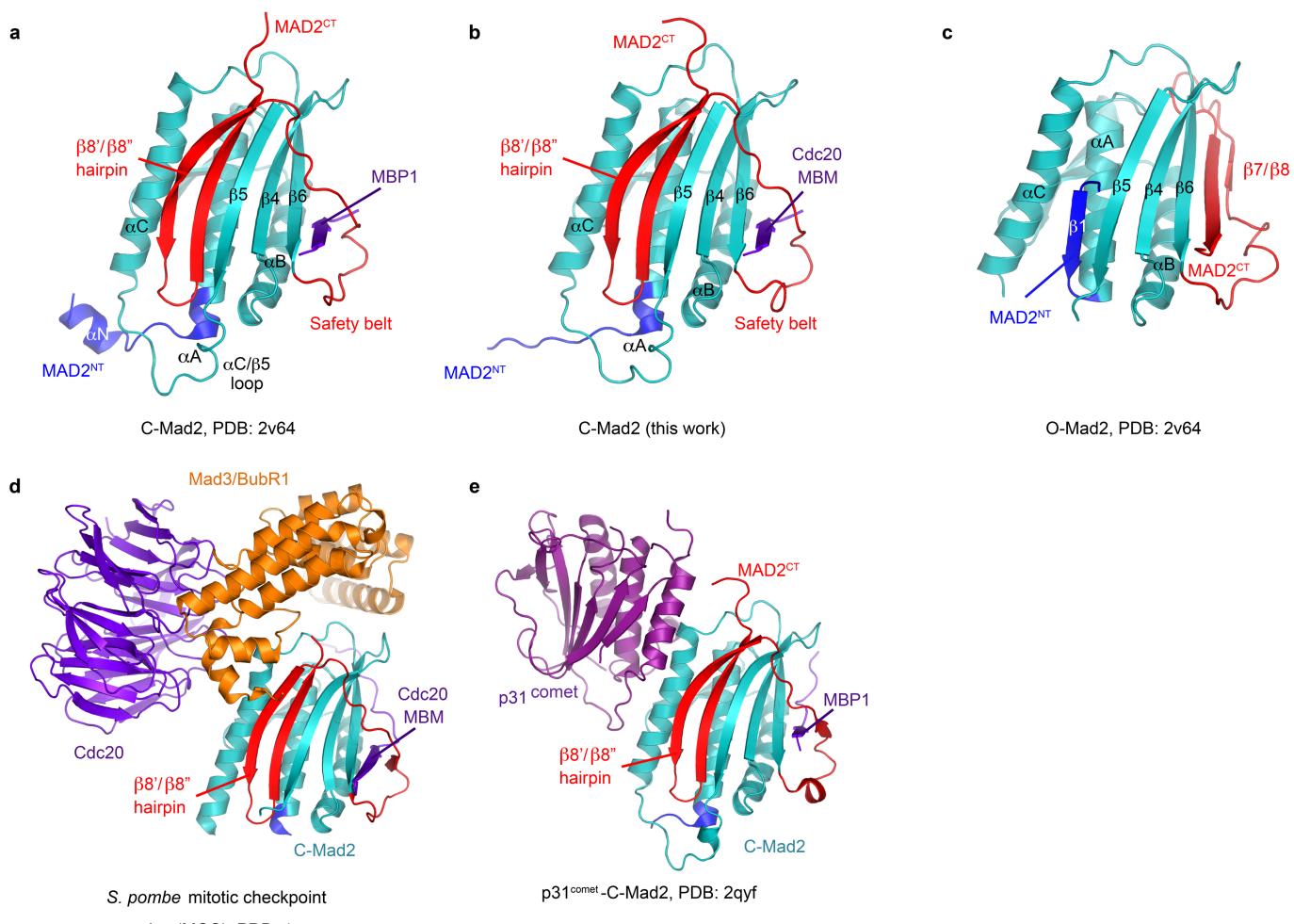
Extended Data Fig. 4 | 3D classification of TRIP13–p31–substrate full data set. **a**, Local refinement (see Methods) by applying a mask covering $\text{TRIP13}^{\text{A/B/C/D/E}}$ monomers. **b, c**, 3D class averages obtained by

classification (see Methods) of the TRIP13–p31–substrate full data set **(b)** and TRIP13–p31–substrate **(c)**. The percentages relative to the total number of TRIP13–p31–substrate particles are shown.



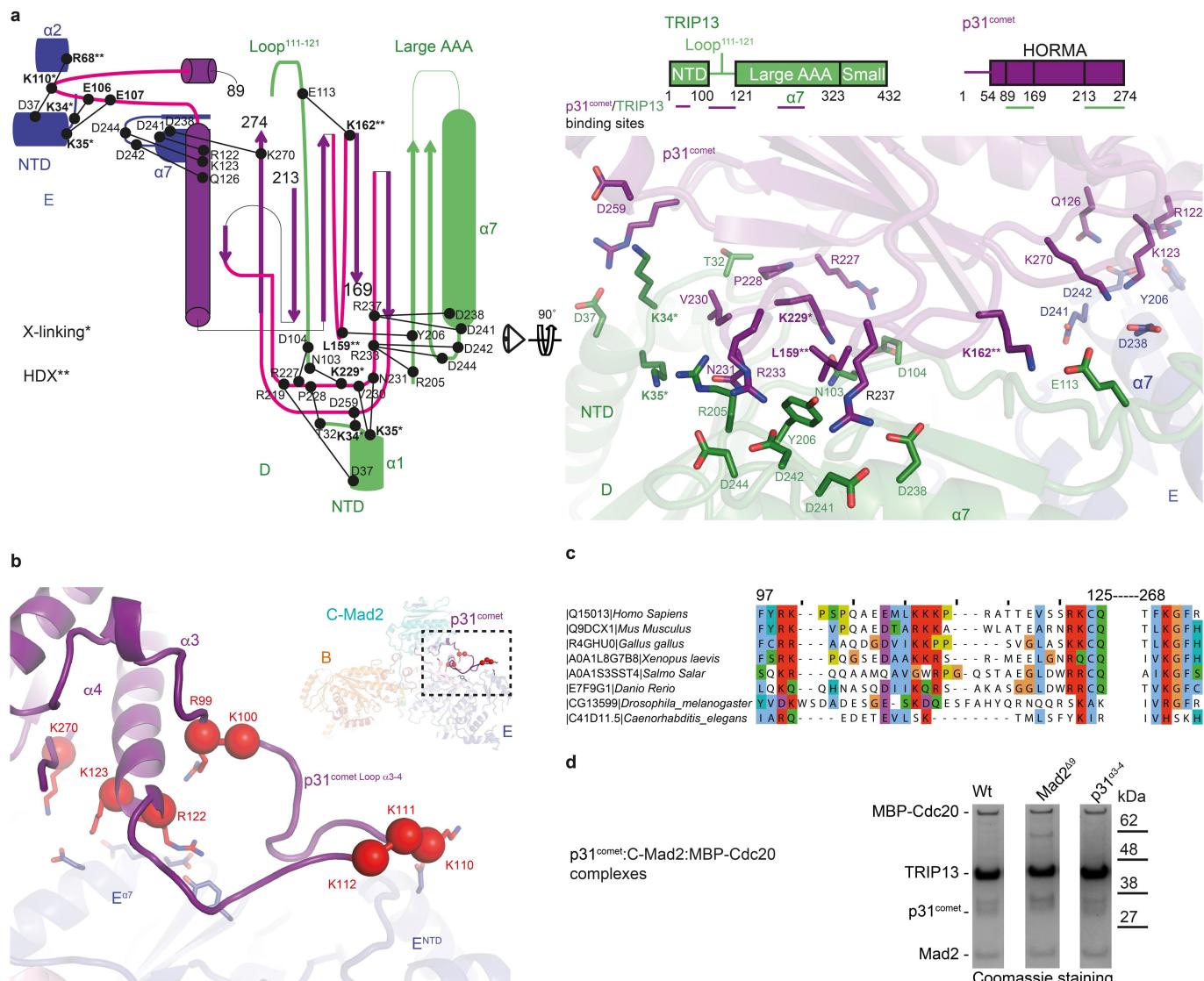
Extended Data Fig. 5 | Comparative analysis of TRIP13 structures. **a**, Comparison of TRIP13 cryo-EM structure (right, this study) and previous TRIP13 crystal structures (left, *C. elegans* PCH2¹⁷; middle, human TRIP13¹⁸). **b**, Comparison of TRIP13 monomers within the TRIP13 cryo-EM structure. RMSDs between TRIP13^A and other TRIP13 monomers are indicated in the inset table. A superimposition of all six TRIP13 monomers is shown, colour-coded as in Fig. 1. TRIP13^F differs from all the other monomers in the relative orientation of the small and

large AAA+ domains. Its conformation relative to TRIP13^A is shown at the lower right. The open conformation prevents nucleotide binding. (The sensor 2 residue (S2) is positioned too far from the ATP-binding site.) Lower left, a superimposition of TRIP13^F onto an open subunit C (grey) of the PCH2 structure¹⁷. **c**, Conformational differences in pore loop-1 between the PCH2-ADP¹⁷ complex (grey) and the cryo-EM TRIP13-ATP-γ-S complex (orange) (this study).



Extended Data Fig. 6 | Structures of MAD2. Structural context of MAD2. **a**, C-MAD2²³; **b**, in the TRIP13–p31^{comet}–C-MAD2 complex (this work); **c**, O-MAD2²³; **d**, the *Schizosaccharomyces pombe* mitotic checkpoint complex⁴² composed of C-MAD2, CDC20 and BUBR1/MAD3 (BUB3 is not shown); **e**, p31^{comet}–C-MAD2 complex¹⁹. In all figures the regions of MAD2 that reposition during the O-MAD2 to C-MAD2 transition are coloured blue and red for N-terminal (residues 1–16) and C-terminal (158–204) regions, respectively. In O-MAD2 these are the N terminus

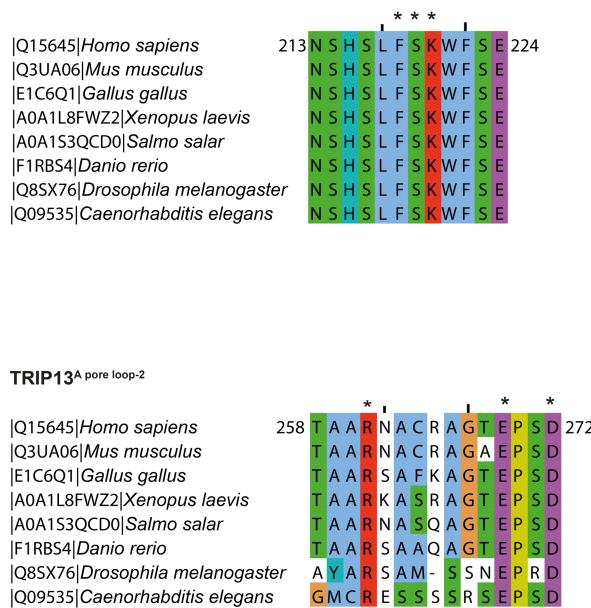
(MAD2^{NT}) and β 1 strand (blue) and C-terminal β 7– β 8 hairpin (red). In C-MAD2 these are MAD2^{NT} including the α N helix and first turn of α A (blue), and the C-terminal β 8'– β 8'' hairpin, safety belt and C terminus (MAD2^{CT}) (red). On conversion of O-MAD2 to C-MAD2, the β 1 strand is displaced and replaced by the β 8'– β 8'' hairpin. Residues 13–15 of β 1 form an additional turn at the N terminus of α A in C-MAD2. The C-MAD2 ligand is coloured purple. MBM, MAD2-binding motif; MBP1, high-affinity MAD2-binding peptide^{43,44}



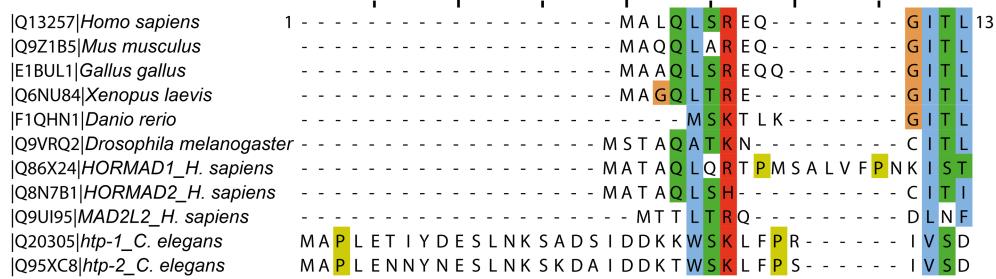
Extended Data Fig. 7 | TRIP13 interacts with p31^{comet} through a composite interface formed of monomers D and E. **a**, Details of the interaction between TRIP13 and p31^{comet}. Left, schematic of TRIP13–p31^{comet} interactions. Right, structure showing details of the main electrostatic contacts between TRIP13 and p31^{comet}. Above, schematic of the domain architecture of TRIP13 and p31^{comet}. A row of aspartates on $\alpha 7$ engages the conserved safety belt motif residues Arg233 and Arg237 of p31^{comet}. The adjacent Lys162 contacts Glu-rich loop (111–121) in TRIP13 that is disordered in previous TRIP13 crystal structures^{17,18}. In our structure the Glu-rich loop lies directly above pore loop-1. Glu104 and Asp105 of the TRIP13^{NTD}-ATPase domain linker, immediately preceding the Glu-rich loop, contact Arg227 and Lys229 of the p31^{comet} safety-belt, agreeing with the importance of Lys229 for TRIP13-p31^{comet} interactions in vitro¹⁷ and in vivo²⁰. On monomer E, the same acidic patch of $\alpha 7$ of

the large AAA+ domain contacts basic residues at the N terminus of $\alpha 3$ of p31^{comet}. **b**, Details of the interaction of the p31^{comet} $\alpha 3$ –4 loop with TRIP13 subunit E. Seven basic residues shown were deleted and the mutant p31^{comet}($\alpha 3$ –4 loop) was tested in MAD2 remodelling assays and for assembly of a TRIP13–p31^{comet}–C-MAD2 complex. **c**, Multiple sequence alignment of the p31^{comet} $\alpha 3$ –4 loop. **d**, Deletion of the nine N-terminal residues of MAD2 (MAD2($\Delta 9$)), and mutation of the p31^{comet} $\alpha 3$ –4 loop (p31^{comet}($\alpha 3$ –4 loop)) do not disrupt TRIP13–p31^{comet}–C-MAD2 complex assembly. Coomassie-stained gel showing the gel filtration fraction of wild-type and relevant mutant TRIP13–p31^{comet}–C-MAD2 complexes purified by size exclusion chromatography. Experiment in **d** was performed in triplicate with similar results. See Supplementary Fig. 1 for gel source data.

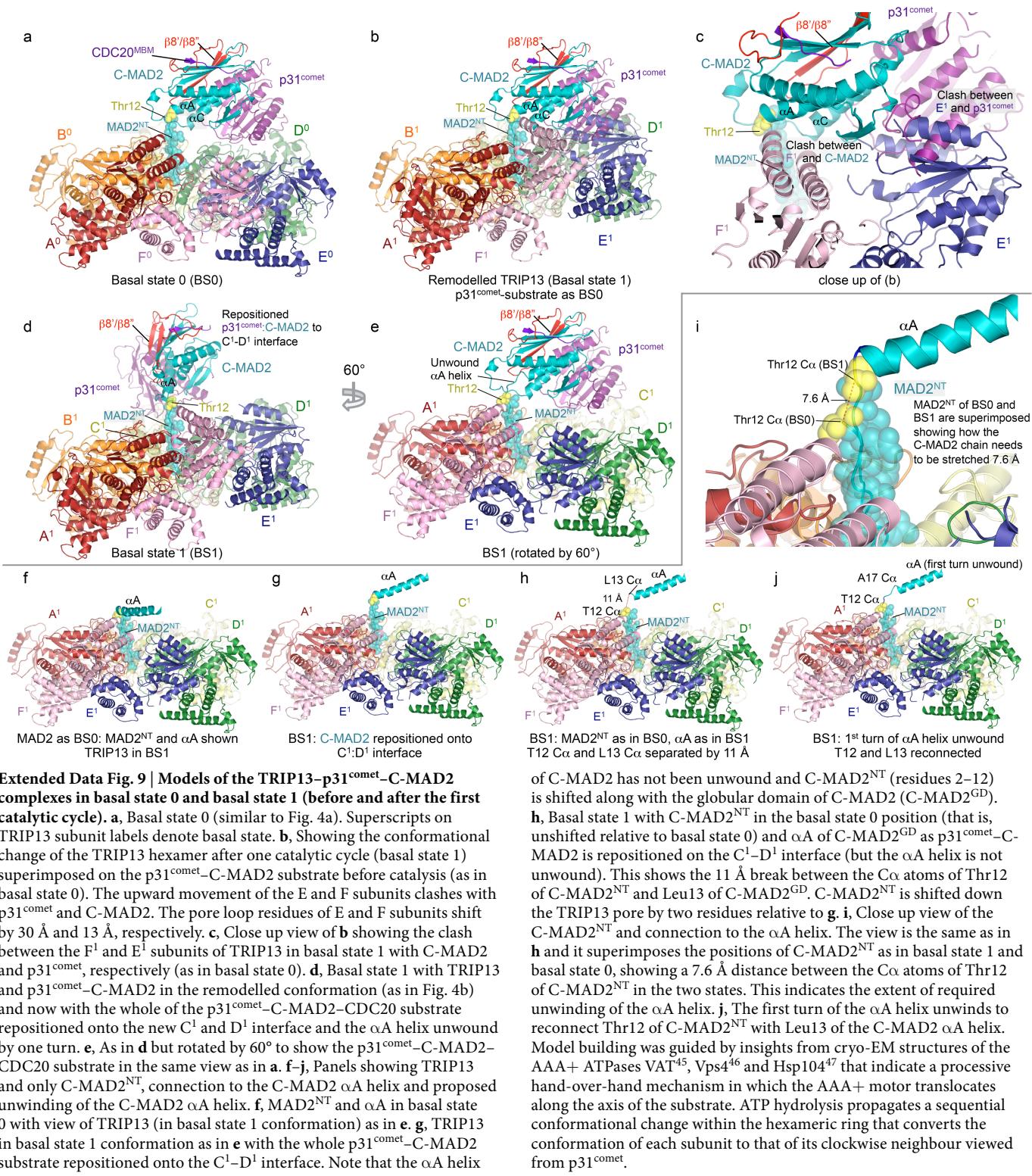
a TRIP13 A pore loop-1



C Mad2^{NT}



Extended Data Fig. 8 | Conservation of TRIP13 pore loops and MAD2^{NT}. a–c, Multiple sequence alignment of TRIP13 pore loop-1 (a) and pore loop-2 (b) and the N-terminal region of MAD2 (c).



Extended Data Table 1 | EM data collection, processing statistics and structure refinement statistics

TRIP13 in complex with ATP-gamma-S, p31comet, C-Mad2 and Cdc20 (EMDB-4166) (PDB 6F0X)	
Data collection and processing	
Magnification	81,000
Voltage (kV)	300
Electron exposure (e ⁻ /Å ²)	40
Defocus range (μm)	2.0-3.6
Pixel size (Å)	1.43
Symmetry imposed	C1
Initial particle images (no.)	440,853
Final particle images (no.)	354,157
Map resolution (Å)	4.6
FSC threshold	0.143
Map resolution range (Å)	4.2-5.5
Refinement	
Initial model used (PDB code)	5vqa, 2qyf
Model resolution (Å)	4.2
FSC threshold	-
Model resolution range (Å)	-
Map sharpening <i>B</i> factor (Å ²)	-342.45
Model composition	
Non-hydrogen atoms	20491
Protein residues	2541
Ligands	5
<i>B</i> factors (Å ²)	
Protein	-
Ligand	-
R.m.s. deviations	
Bond lengths (Å)	0.01
Bond angles (°)	1.13
Validation	
MolProbit score	2.62
Clashscore	27.73
Poor rotamers (%)	0
Ramachandran plot	
Favored (%)	79.26
Allowed (%)	20.19
Disallowed (%)	0.55

High speed of fork progression induces DNA replication stress and genomic instability

Apolinar Maya-Mendoza^{1,4*}, Pavel Moudry^{1,2,4}, Joanna Maria Merchut-Maya¹, MyungHee Lee¹, Robert Strauss¹ & Jiri Bartek^{1,2,3*}

Accurate replication of DNA requires stringent regulation to ensure genome integrity. In human cells, thousands of origins of replication are coordinately activated during S phase, and the velocity of replication forks is adjusted to fully replicate DNA in pace with the cell cycle¹. Replication stress induces fork stalling and fuels genome instability². The mechanistic basis of replication stress remains poorly understood despite its emerging role in promoting cancer². Here we show that inhibition of poly(ADP-ribose) polymerase (PARP) increases the speed of fork elongation and does not cause fork stalling, which is in contrast to the accepted model in which inhibitors of PARP induce fork stalling and collapse³. Aberrant acceleration of fork progression by 40% above the normal velocity leads to DNA damage. Depletion of the treslin or MTBP proteins, which are involved in origin firing, also increases fork speed above the tolerated threshold, and induces the DNA damage response pathway. Mechanistically, we show that poly(ADP-ribosylation) (PARylation) and the PCNA interactor p21^{Cip1} (p21) are crucial modulators of fork progression. PARylation and p21 act as suppressors of fork speed in a coordinated regulatory network that is orchestrated by the PARP1 and p53 proteins. Moreover, at the fork level, PARylation acts as a sensor of replication stress. During PARP inhibition, DNA lesions that induce fork arrest and are normally resolved or repaired remain unrecognized by the replication machinery. Conceptually, our results show that accelerated replication fork progression represents a general mechanism that triggers replication stress and the DNA damage response. Our findings contribute to a better understanding of the mechanism of fork speed control, with implications for genomic (in)stability and rational cancer treatment.

Single-stranded DNA (ssDNA) breaks activate PARP1, which PARylates proteins to regulate diverse cellular processes⁴. Inhibition of PARP1 is toxic in tumours deficient in *BRCA1* and *BRCA2* genes, presumably owing to impaired replication fork protection⁵ and persistence of unrepaired collapsed forks^{3,6}. PARP inhibitors, such as olaparib, have been approved for clinical use in the treatment of cancer⁷. Although a role of PARylation in collapsed fork recovery has been suggested⁸, the effect of PARP inhibitors during unperturbed S phase remains unclear. Increasing doses of olaparib induced the accumulation of cells in S/G2 and inhibited cell proliferation (Extended Data Fig. 1a–c). DNA fibre analysis showed that treatment with 10 μ M olaparib for 24 h increased the speed of fork progression by 60% (Fig. 1a, b). Fork acceleration was dependent on the dose and timing of the PARP inhibitor (PARPi) (Extended Data Fig. 1d–g) but independent of cell type (Extended Data Fig. 1h, i). To identify stalled and collapsed forks, we analysed symmetry between the first and second pulse in double-labelled DNA fibres⁹. Unexpectedly, treatment with olaparib did not stall forks (Fig. 1c). Veliparib, another PARPi, also accelerated forks without affecting fork symmetry and caused S/G2 accumulation (Extended Data Fig. 1j–l). Analysis of the replication program showed that olaparib-treated cells accumulated in mid-to-late S phase (Fig. 1d; Extended Data Fig. 2), and activated hallmarks of

the DNA damage response (DDR) including foci of γ H2AX, RAD51, RPA, 53BP1 and phosphorylation of RPA and CHK1 (Extended Data Fig. 3a–c). Among the effects of PARPi-induced replication stress, PARP-inhibited cells showed increased tail moment in alkaline but not neutral comet assays (Fig. 1e, f; Extended Data Fig. 3d). Analysis by TUNEL assay in PARPi-treated cells showed foci located next to PCNA-containing replication machineries (Fig. 1g, h; Extended Data Fig. 3e), indicating the generation of DNA nicks or breaks during or immediately after DNA replication. Quantification of ssDNA in PARP-inhibited cells by detecting incorporated BrdU under non-denaturing conditions revealed increased numbers of cyclin-A-positive cells with more than five BrdU foci (Fig. 1i). PARP-inhibited cells also incorporated EdU in G2 phase (Fig. 1j, k), accumulated mitotic chromosome bridges (Fig. 1l, m) and formed micronuclei (Fig. 1n). Together, our results indicate that the lack of PARylation induces replication stress, accelerates fork speed without stalling forks and triggers the DDR pathway.

To test our hypothesis that fast fork progression induces DDR, we altered the abundance of proteins involved in origin firing, because the number of active origins influences fork speed¹⁰. Treslin-knockdown cells featured faster forks (around 1.5 kb min⁻¹; Fig. 2a, Extended Data Fig. 4a, b), activated DDR, accumulated in late S phase, and moderately increased asymmetric forks (Fig. 2b–e). Knockdown of MDM2-binding protein (MTBP) also accelerated fork extension (around 1.4 kb min⁻¹) and triggered DDR (Fig. 2a, b; Extended Data Fig. 4c–e), whereas knockdown of TopBP1 resulted in slow and asymmetric fork progression (Fig. 2a, b). Combined knockdown of treslin and PARPi did not further accelerate fork extension (both 1.7 kb min⁻¹), but it did inhibit proliferation, with accumulation of S-phase cells (Extended Data Fig. 4f–h). PARPi treatment did not alter the levels of treslin protein (Extended Data Fig. 4g). Thus, the effects of PARPi-triggered fork acceleration do not operate through deregulation of treslin, and fork acceleration generated by two means—PARPi or imbalance of proteins involved in origin firing—share the ability to trigger DDR.

ATR inhibition (ATRi) triggers maximum origin activity¹¹. To estimate origin firing, we measured the distances between origins of replication (Fig. 2f; Extended Data Fig. 4i) and counted fork density per 1 Mb of DNA (Extended Data Fig. 4j, k). An average origin-to-origin distance was 130.9 kb in untreated (non-targeting control) cells, 218 kb in PARP-inhibited cells, 35 kb in ATR-inhibited cells, and 38 kb in PARP/ATR-inhibited cells (Fig. 2f). Treslin knockdown increased the inter-origin distance to 195 kb. Consistent with the role of treslin in origin regulation, ATRi in treslin-knockdown cells did not fully increase the number of origins (treslin-knockdown/ATRi = 89.8 kb versus ATRi = 35 kb, $P = 4 \times 10^{-36}$). As fork speed affects origin usage¹⁰, we examined whether fork extension is affected by ATRi in PARP-inhibited cells. ATRi decreased fork speed in mock-treated (0.3 kb min⁻¹) and PARPi/ATRi-treated (0.3 kb min⁻¹) cells. Owing to the inability to fire the full set of origins, treslin knockdown plus ATRi treatment did not reduce fork speed (treslin-knockdown/ATRi = 0.9 kb min⁻¹; ATRi = 0.3 kb min⁻¹; Fig. 2g, Extended Data

¹Genome Integrity Unit, Danish Cancer Society Research Center, Copenhagen, Denmark. ²Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czech Republic. ³Division of Genome Biology, Department of Medical Biochemistry and Biophysics, Science for Life Laboratory, Karolinska Institute, Stockholm, Sweden. ⁴These authors contributed equally: Apolinar Maya-Mendoza, Pavel Moudry. *e-mail: apomm@cancer.dk; jb@cancer.dk

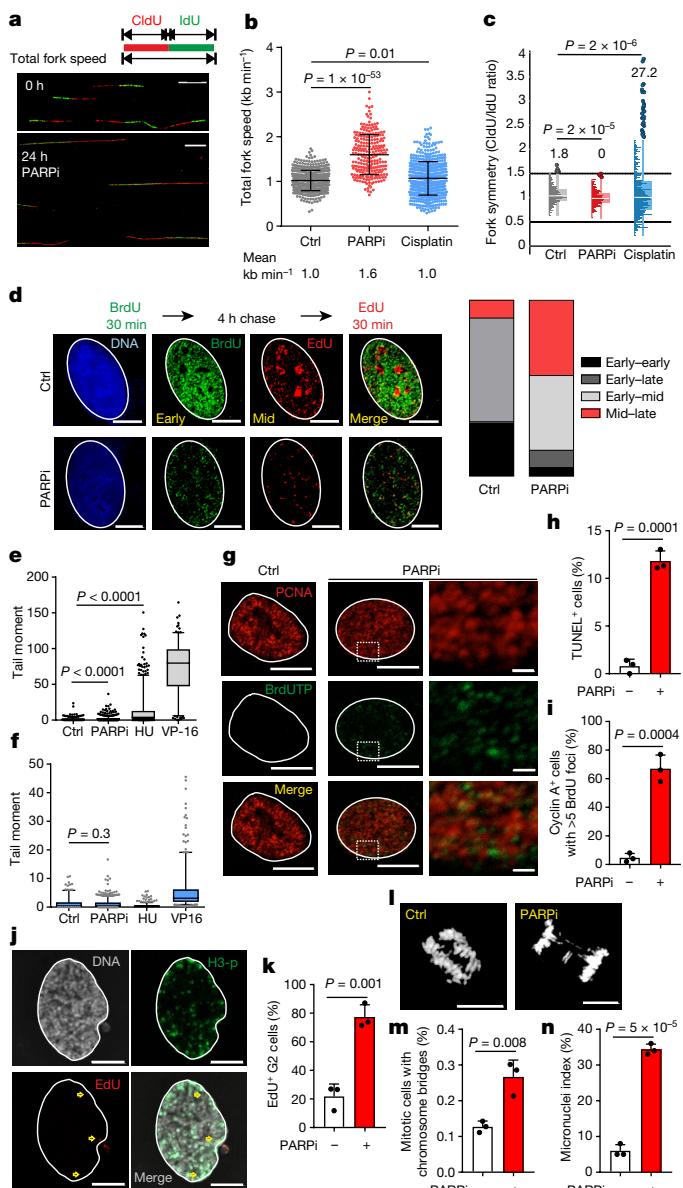


Fig. 1 | PARP inhibition induces fork acceleration and replication stress. **a, b**, DNA fibres from U2-OS cells treated with 10 μ M of the PARPi olaparib or 30 μ M cisplatin for 24 h. Mean fork speed (kb min^{-1}) is indicated. Scored forks: control (ctrl) = 503; PARPi = 244; cisplatin = 552. **c**, CldU/IdU ratios from values in **b**. The percentage of highly asymmetric forks (CldU/IdU ratios < 0.5 and > 1.5) is indicated. **d**, S-phase progression in U2-OS cells; $n > 250$ nuclei. **e, f**, Alkaline (**e**) and neutral (**f**) comet assays from U2-OS cells treated with PARPi 10 μ M for 24 h, 2 mM hydroxyurea (HU), or 10 μ M VP-16. Hydroxyurea and VP-16 are positive controls for ssDNA and dsDNA, respectively. **g**, TUNEL assay in U2-OS cells. **h**, TUNEL-positive cells from Extended Data Fig. 3e. **i**, BrdU incorporation under non-denaturing conditions in cyclin-A-positive cells (BrdU added for 48 h and PARPi for the last 24 h). **j, k**, EdU (arrows) and phospho-S10 histone 3 (H3-p) in U2-OS cells. **l**, Normal mitosis (control) or mitotic chromosome bridges (PARPi). **m**, Anaphases scored PARPi (−) $n = 176$; PARPi (+) $n = 180$. **n**, Cells with micronuclei after 3-day treatment with PARPi. Scale bars, 10 μ m and 1 μ m (expanded magnified region to the right in **g**). Data are mean \pm s.d.; for statistics and reproducibility, see accompanying Source Data.

Fig. 4l). Thus, PARP-inhibited cells can potentially activate the full repertoire of replication origins.

A conceptually important question was whether there is a threshold of fork speed beyond which cells trigger DDR. Given that depletion of other replication-regulating proteins accelerates fork progression¹²,

we depleted ligase 1 (LIG1) and Flap endonuclease 1 (FEN1). These manipulations accelerated fork speed to approximately 1.3 and 1.2 kb min^{-1} , respectively, without affecting the cell cycle or activating DDR (Fig. 2h, i; Extended Data Fig. 4m). Titration of fork speed by using a low concentration of PARPi or a short incubation time (Extended Data Fig. 1d, f) resulted in fork acceleration below 40%, and cell cycle and cell viability remained unaffected (Extended Data Fig. 5). These results suggest that cells can buffer changes in fork speed, but if velocity increases above a 40% gain, genomic integrity is compromised (Fig. 2j). The possibility that fork acceleration after PARPi reflects altered chromatin structure was then tested and excluded, as no PARPi-induced changes in global chromatin configuration were detected (Extended Data Fig. 6).

PARP inhibition versus PARP1 depletion lead to different outcomes in terms of ssDNA break repair¹³. In our experiments, PARP1 knockdown slightly affected the fork rate (1.0 versus 1.2 kb min^{-1} , Fig. 3a) and shifted the cell cycle towards G1 enrichment (Fig. 3h, Extended Data Fig. 9b). PARPi in PARP1-depleted cells did not further accelerate fork speed (PARP1 knockdown = 1.2 kb min^{-1} ; PARP1 knockdown/PARPi = 1.2 kb min^{-1} , $P = 0.56$). If fast fork extension causes DDR, then treatment of PARP1-knockdown cells with PARPi should not trigger DDR. Indeed, PARP1 knockdown alleviated the DDR-triggering effect of PARPi (Fig. 3b, c; Extended Data Fig. 7a). Therefore, the effect of PARPi on fork elongation speed requires the PARP1 protein. The basal nuclear PARylation level in PARP1-knockdown cells was unchanged (Fig. 3d; Extended Data Fig. 7b, c), suggesting redundancy by other PARP protein(s), and providing motivation to search for additional PARP-related mechanism(s) of fork speed control.

p53 transactivates p21 (also known as CDKN1A)¹⁴, p21 can inhibit DNA synthesis by binding to PCNA¹⁵, and PARylation regulates p53 activity¹⁶. PARP1 physically binds to p21¹⁷. Moreover, PARP1 is a co-repressor of the p21 gene promoter¹⁸. By exploring the potential interaction between PARP1, PARylation and the p53–p21 pathway in the regulation of fork speed, we observed that PARP1 knockdown resulted in increased p21 abundance (Fig. 3e; Extended Data Fig. 7d). Depletion of p21 using short interfering RNAs (siRNAs) or short hairpin RNA (shRNA) accelerated fork elongation without affecting the PARylation level. Conversely, PARPi slightly decreased the nuclear p21 level (Fig. 3f, g; Extended Data Fig. 7d–h). We also observed an additive effect of fork acceleration in p21-knockdown cells treated with PARPi (2.1 kb min^{-1} , Fig. 3f). Cell cycle alterations seen in PARP1-depleted cells were alleviated in double PARP1 and p21 knockdown cells, consistent with close-to-normal fork speed (Fig. 3f, h). Knockdown of p53 did not alter the nuclear PARylation (Extended Data Fig. 7i) and slowed down fork progression (0.8 kb min^{-1} , Fig. 3f), probably reflecting pleiotropic effects of p53 absence, including enhanced DNA damage (Fig. 3g) and loss of fork stability¹⁹. Our p21-depletion experiments pinpoint p21 as a negative regulator of fork speed, consistently with reports that p21 inhibits DNA synthesis^{15,20,21}, but in contradiction to Mansilla et al.²², whose results on fork speed we could not reproduce (Extended Data Fig. 7j, k).

Our results using U2-OS cells show that PARP1-knockdown induced p21 expression without altering PARylation, indicating that in cells with the functional p53–p21 axis, and with other PARP proteins compensating for PARP1, forks could not elongate as fast as in PARP-inhibited cells. We predicted that in cells that lack PARP2 and with disabled p53–p21, forks should elongate fast even after PARP1 knockdown, because PARylation might not be maintained and p21 may remain low. Indeed, in HeLa cells that express PARP2 and p21 poorly (Extended Data Fig. 8a; <http://www.proteinatlas.org>)²³, the knockdown of PARP1 resulted in low PARylation, no p21 upregulation, and fork speed accelerated similarly in the PARP-inhibited and PARP1-knockdown cells (Fig. 3i; Extended Data Fig. 8b–e). Furthermore, double knockdown of PARP1 and PARP2 in U2-OS cells neither accelerated fork speed (non-targeting knockdown control = 0.94 kb min^{-1} ;

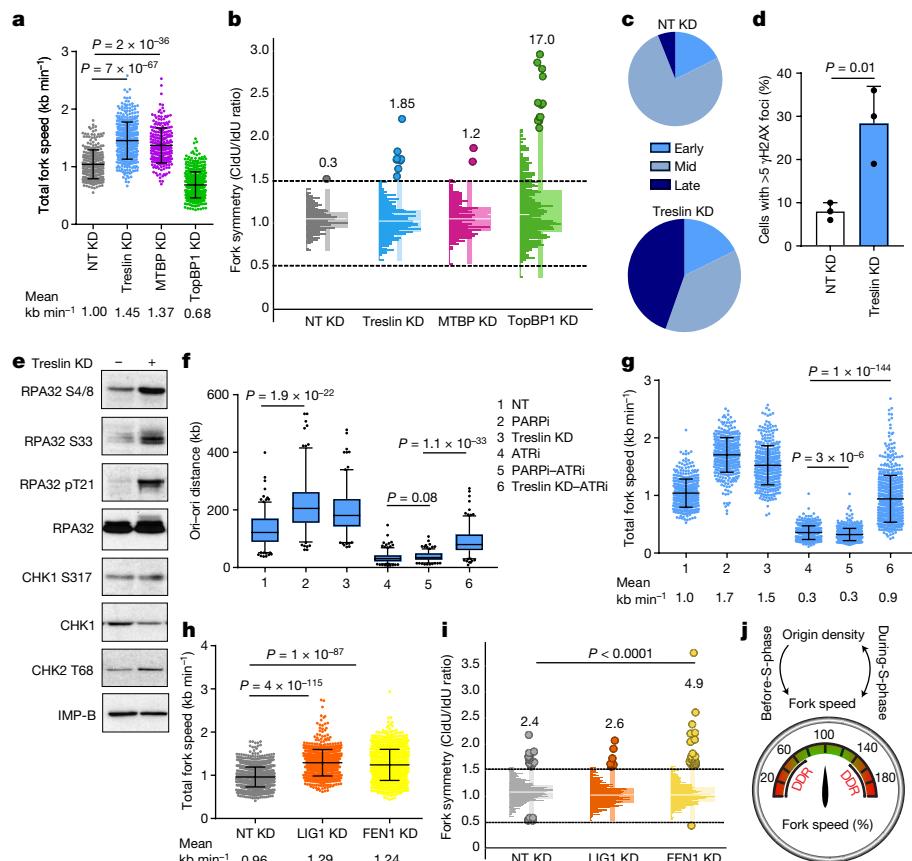


Fig. 2 | Fork speed threshold and DDR. **a**, DNA fibres from non-targeting (NT) knockdown control, or treslin-, MTBP- or TopBP1-knockdown (KD) U2-OS cells, 72 h after transfection. Scored forks: NT = 293; treslin KD = 431; MTBP KD = 250; TopBP1 KD = 494. **b**, CldU/IdU ratios from values in **a**. Total fork speed = $(\text{CldU} + \text{IdU}) \text{ t}^{-1}$. **c**, Replication program by BrdU incorporation in non-targeting or treslin-knockdown U2-OS cells. **d**, Percentage of non-targeting or treslin-knockdown U2-OS cells with more than five γ H2AX foci. **e**, Immunoblots of DDR proteins in non-targeting or treslin-knockdown U2-OS cells. IMP-B is a loading control. **f**, Origin-to-origin distance (kb) in non-targeting or treslin-knockdown U2-OS cells (10 μ M PARPi, 24 h; 1 μ M ATRi, 1 h). Mean origin-to-origin distance: NT = 130.9 kb; PARPi = 218.6 kb; treslin KD = 195.6 kb;

ATRi = 35.34 kb; PARPi-ATRi = 38.23 kb; treslin KD-ATRi = 89.84 kb. Whiskers indicate the fifth and ninety-fifth percentiles, and the centre values depict the median. **g**, DNA fibres from non-targeting or treslin-knockdown U2-OS cells (treatment as in **f**). Scored forks: NT = 410; PARPi = 395; treslin KD = 398; ATRi = 439; PARPi-ATRi = 520; treslin KD-ATRi = 580. **h**, DNA fibres from non-targeting, LIG1- or FEN1-knockdown U2-OS cells. Scored forks: NT = 1362; LIG1 LD = 739; FEN1 KD = 960. **i**, CldU/IdU ratios from values in **h**. **j**, Fork speed that is increased above 40% or reduced by at least 20% induces DDR (red zone). Cells buffer some changes in fork speed without triggering DDR (green zone). Data are mean \pm s.d.; for statistics and reproducibility, see accompanying Source Data.

PARP1/2-knockdown = 1.06 kb min^{-1}) nor affected the cell cycle. Moreover, PARPi-treated PARP1/2-knockdown cells were unable to highly increase fork velocity (non-targeting/PARPi = 1.7 kb min^{-1} ; PARP1/2-knockdown/PARPi = 1.26 kb min^{-1}), consistent with increased p21 compared with untreated cells (Extended Data Fig. 9). These results support the notion that PARylation and p21 act as key regulators of replication fork speed.

Relevant to the clinically exploited synthetic lethality of BRCA1 dysfunction that sensitizes cells to PARPi⁷, we found that in BRCA1-knockdown U2-OS cells treated with PARPi, forks accelerated to 1.6 kb min^{-1} (Fig. 4a), accompanied by some fork asymmetry (Fig. 4b). Notably, fork asymmetry was reduced, rather than increased, in BRCA1-knockdown cells treated with PARPi compared to control BRCA1-knockdown cells, suggesting that PARylation is required to detect damaged DNA and arrest forks. Moreover, in BRCA1-knockdown cells, PARPi increased the activity of ATM and ATR kinases (Fig. 4c). Although defective forks that occur in control BRCA1-proficient cells stall, in BRCA1-deficient/PARP-inhibited cells, the ability of forks to stall is impaired. Consistently, fork speed in BRCA1-deficient breast cancer cells MDA-MB-436 was accelerated by PARPi, and the number of stalled forks was reduced (Fig. 4d, e). These results suggest that synthetic lethality seen in BRCA1-deficient cells treated with PARPi might not reflect acute fork stalling and ensuing

accumulation of collapsed forks, as generally assumed^{3,6}, but could instead reflect the initial acceleration of fork progression beyond the tolerable threshold and accumulation of unrecognized DNA damage, eventually affecting cell viability (Extended Data Fig. 10a, b). As PARP inhibitors have been approved for the treatment of ovarian cancer, we also treated human ovarian cancer cells (OVCAR-5) with olaparib, and again found accelerated fork elongation affecting the cell cycle and viability (Fig. 4f, g; Extended Data Fig. 10c, d). Furthermore, we found inability to recognize DNA damage that potentially affects fork progression in TopBP1-depleted PARPi-treated cells (Fig. 4h, i), probably contributing to synthetic lethality observed in TopBP1-deficient cells under PARPi²⁴.

Arrested forks were not observed under PARPi treatment, indicating that a 'sensor' of replication stress was disabled when PARylation was inhibited. Indeed, p21-knockdown caused accumulation of some asymmetric forks, which were absent under PARP inhibition (Fig. 4j). Therefore, PARylation provides a sensor of replication stress at active replication forks, as PARP1 activation allows replication stress signalling and can lead to fork arrest. Moreover, activated PARP1 targets p53 and causes p53-mediated accumulation of p21 that can then stop defective forks from further progression.

We suggest a model in which PARylation and the p53-p21 axis provide a coordinated mechanism that we call a fork speed regulatory network

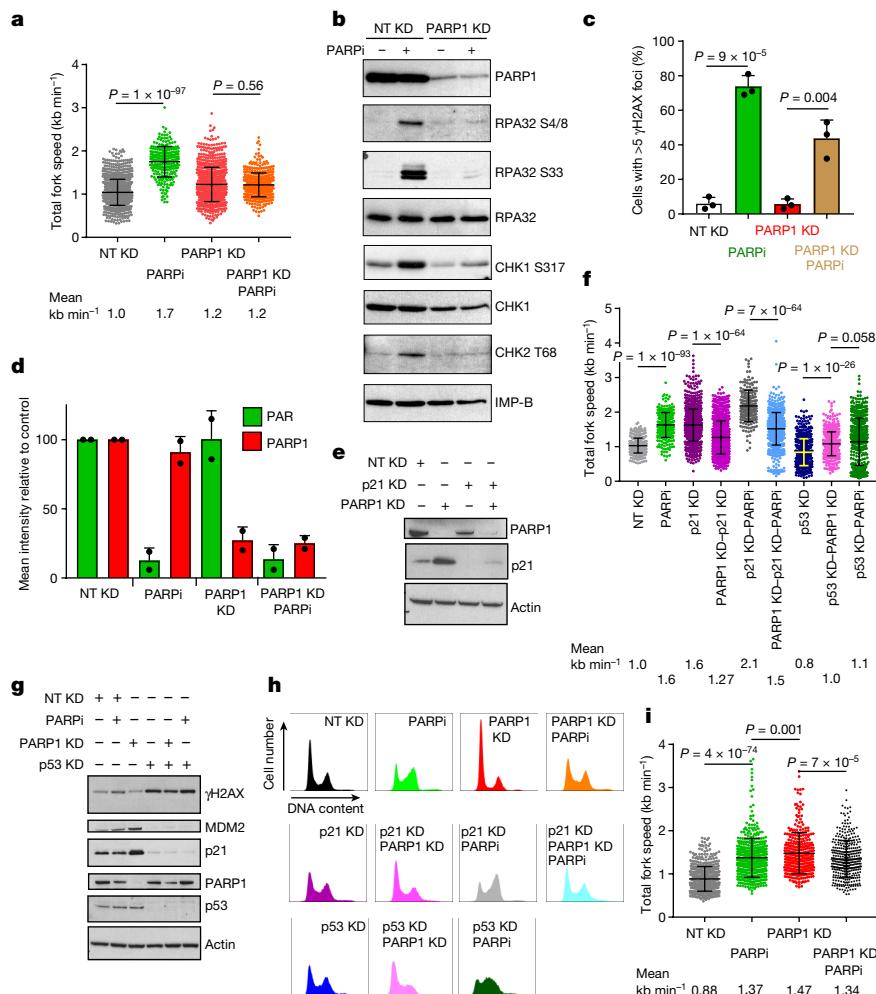


Fig. 3 | PARP1, its activity and p21 regulate fork speed. **a**, DNA fibres from non-targeting control or PARP1-knockdown U2-OS cells, 72 h after transfection (10 μ M PARPi, 24 h). Scored forks: NT = 1,179; PARPi = 255; PARP1 KD = 855; PARP1 KD-PARPi = 453. **b**, Immunoblots of DDR proteins in non-targeting or PARP1-knockdown U2-OS cells (lines 2, 4 as in **a**). **c**, Percentage of non-targeting or PARP1-knockdown U2-OS cells with more than five γ H2AX foci. **d**, Mean intensity of poly(ADP-ribose) (PAR) and PARP1 in non-targeting or PARP1-knockdown U2-OS cells (treatment as in **a**). **e**, Immunoblots of PARP1 and p21 in PARP1-knockdown and/or p21-knockdown U2-OS cells. Actin is a loading control. **f**, DNA fibres from non-targeting, p21-knockdown, p53-knockdown, double PARP1 and p21 knockdown, or double p53

(FSRN), which negatively regulates fork speed in S phase. The two branches of this network also show mutual feedback links (Fig. 4k). Our dataset complements the concept of PARP proteins being trapped at the site of DNA damage after PARP inhibition²⁵. These potential replication obstacles might be ignored owing to the lack of PARP activity and/or impaired p53-regulatory signalling. Thus, without p53 and/or PARylation, cell fitness is severely compromised. Another aspect of our present results is the positive role of PARP activity in recruitment of DDR proteins to defective forks. Finally, PARPi may impair fork reversal, an important mechanism to protect forks from breakage²⁶.

The concerted actions of PARP1 activity, p53 and p21 are crucial to maintain correct speed and fidelity of DNA synthesis (Fig. 4k). Within the FSRN, activated PARP1 may interact directly with the replication machinery to respond to any fork blockage promptly (Extended Data Fig. 10e). The PARP1-p21 interaction is further complemented by ATR-regulated signalling that is triggered by stretches of ssDNA coated with RPA as a result of uncoupling DNA helicase and polymerase activities in arrested forks¹¹.

and PARP1 knockdown U2-OS cells (treatment as in **a**). Scored forks: NT = 389; PARPi = 314; p21 KD = 974; PARP1 KD-p21 KD = 1,222; p21 KD-PARPi = 210; PARP1 KD-p21 KD-PARPi = 758; p53 KD = 581; p53 KD-PARPi KD = 504; p53 KD-PARPi = 611. **g**, Immunoblots of γ H2AX, MDM2, p21, PARP1 and p53 in PARP1-knockdown, p53-knockdown or double p53 and PARP1 knockdown U2-OS cells (lines 2, 6 as in **a**). **h**, Cell cycle profiles of non-targeting, PARP1-, p21- or p53-knockdown U2-OS cells (treatment as in **a**). **i**, DNA fibres from non-targeting or PARP1-knockdown HeLa cells (treatment as in **a**). Scored forks: NT = 553; PARPi = 473; PARP1 KD = 334; PARP1 KD-PARPi = 378. Data are mean \pm s.d.; for statistics and reproducibility, see accompanying Source Data.

Our data provide an unorthodox rationale for the reported synergistic effect of therapies combined with PARPi. We propose that: (1) PARP inhibition causes replication stress, accelerates fork elongation and induces DDR; (2) replication stress triggered by defective fork recovery and/or homologous recombination-mediated repair might be ignored by cells that lack PARylation²⁷; and (3) PARP inhibitors affect the normal regulation of p53 and its downstream effectors. With this rationale, PARP inhibitors would enhance deleterious effects caused by any insults that negatively affect fork progression. Moreover, increased fork speed may diminish the fidelity of DNA polymerases, potentially contributing to genomic instability²⁸. We provide evidence that PARP inhibitors induce replication stress, simultaneously making cells unresponsive to fork defects, properties beneficial in both monotherapy of highly proliferative cancers and when combined with fork-damaging chemotherapy. Supra-threshold fork speed causing DDR is probably applicable also to (patho-) physiological depletion of proteins involved in origin firing, exemplified by deficit in MCM helicase components resulting in replication stress of old mouse haematopoietic stem cells²⁹, further supporting the relevance of our present concept to both cancer and ageing.

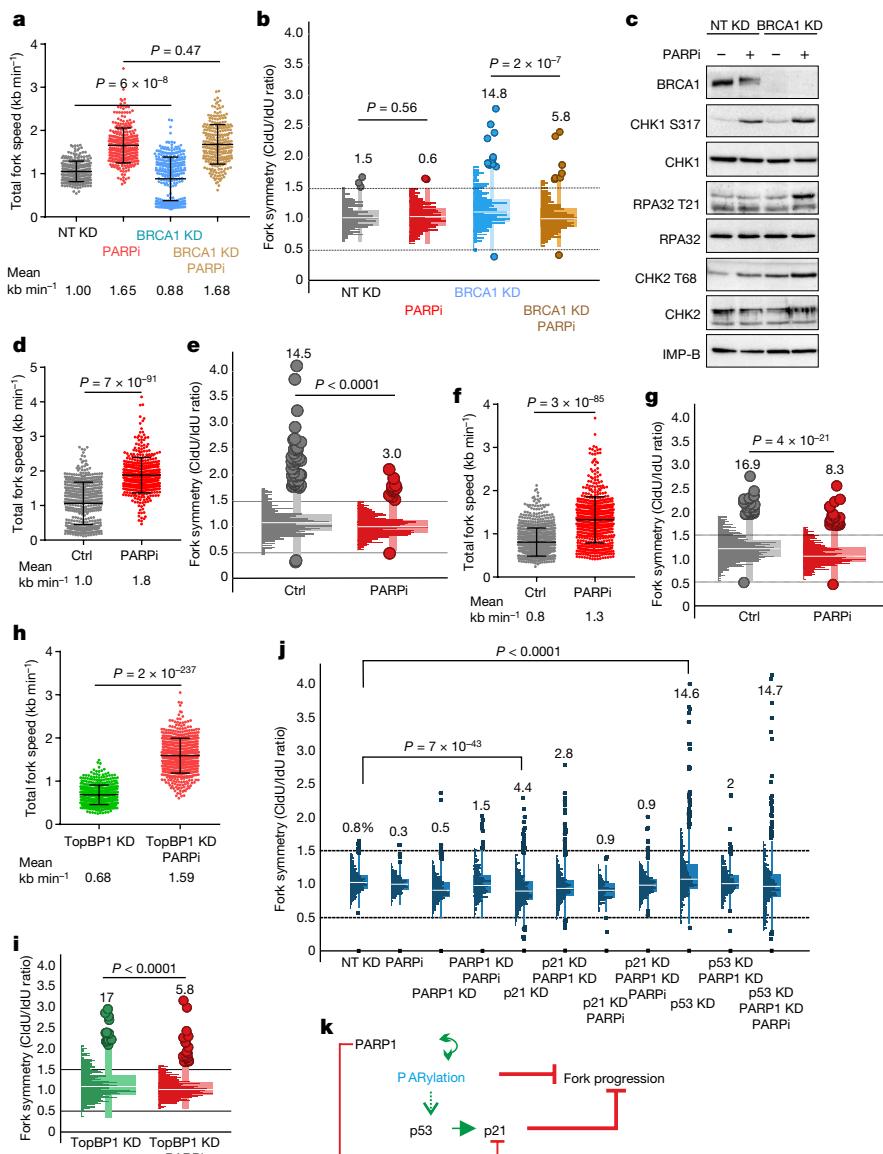


Fig. 4 | Fork stalling is impaired in PARP-inhibited cells. **a**, DNA fibres from non-targeting control or BRCA1-knockdown U2-OS cells, 72 h after transfection (10 μ M PARPi, 24 h). Scored forks: NT = 260; PARPi = 309; BRCA1 KD = 337; BRCA1 KD-PARPi = 257. **b**, CldU/IdU ratios from values in **a**. **c**, Immunoblots of DDR proteins in non-targeting or BRCA1-knockdown U2-OS cells (lines 2, 4 as in **a**). **d**, DNA fibres from BRCA1-deficient MDA-MB-436 cells (treatment as in **a**). Scored forks: control = 508; PARPi = 458. **e**, CldU/IdU ratios from values in **d**. **f**, DNA fibres from OVCAR-5 ovarian cancer cells (treatment as in **a**). Scored forks: control = 844; PARPi = 646. **g**, CldU/IdU ratios from values

in **f**, **h**, DNA fibres from TopBP1-knockdown U2-OS cells (treatment as in **a**). Scored forks: TopBP1 KD = 494; TopBP1 KD-PARPi = 562. **i**, CldU/IdU ratios from values in **h**. **j**, CldU/IdU ratios from values in Fig. 3a, f. **k**, FSRN model in normal S phase and after DNA damage. PARP activity is the initial 'sensor' of replication stress-associated DNA damage. PARylation suppresses fork progression, and can modify p53, which transactivates p21, another suppressor of fork elongation. Red arrows: negative regulation, green arrows: activation (Extended Data Fig. 10). For statistics and reproducibility, see accompanying Source Data.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0261-5>.

Received: 24 June 2016; Accepted: 22 May 2018;
Published online: 27 June 2018

- Conti, C. et al. Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol. Biol. Cell* **18**, 3059–3067 (2007).
- Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319**, 1352–1355 (2008).
- Bryant, H. E. et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913–917 (2005).
- Luo, X. & Kraus, W. L. On PAR with PARP: cellular stress signaling through poly(ADP-ribose) and PARP-1. *Genes Dev.* **26**, 417–432 (2012).
- Schlacher, K., Wu, H. & Jasin, M. A distinct replication fork protection pathway connects Fanconi anemia tumor suppressors to RAD51-BRCA1/2. *Cancer Cell* **22**, 106–116 (2012).
- Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).
- Ledermann, J. et al. Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. *Lancet Oncol.* **15**, 852–861 (2014).
- Bryant, H. E. et al. PARP is activated at stalled forks to mediate Mre11-dependent replication restart and recombination. *EMBO J.* **28**, 2601–2615 (2009).
- Burrell, R. A. et al. Replication stress links structural and numerical cancer chromosomal instability. *Nature* **494**, 492–496 (2013).
- Zhong, Y. et al. The level of origin firing inversely affects the rate of replication fork progression. *J. Cell Biol.* **201**, 373–383 (2013).

11. Toledo, L. I. et al. ATR prohibits replication catastrophe by preventing global exhaustion of RPA. *Cell* **155**, 1088–1103 (2013).
12. Duxin, J. P. et al. Okazaki fragment processing-independent role for human Dna2 enzyme during DNA replication. *J. Biol. Chem.* **287**, 21980–21991 (2012).
13. Godon, C. et al. PARP inhibition versus PARP-1 silencing: different outcomes in terms of single-strand break repair and radiation susceptibility. *Nucleic Acids Res.* **36**, 4454–4464 (2008).
14. el-Deiry, W. S. et al. WAF1, a potential mediator of p53 tumor suppression. *Cell* **75**, 817–825 (1993).
15. Waga, S., Hannon, G. J., Beach, D. & Stillman, B. The p21 inhibitor of cyclin-dependent kinases controls DNA replication by interaction with PCNA. *Nature* **369**, 574–578 (1994).
16. Lee, M. H., Na, H., Kim, E. J., Lee, H. W. & Lee, M. O. Poly(ADP-ribosylation) of p53 induces gene-specific transcriptional repression of MTA1. *Oncogene* **31**, 5099–5107 (2012).
17. Frouin, I. et al. Human proliferating cell nuclear antigen, poly(ADP-ribose) polymerase-1, and p21^{Waf1/Cip1}. A dynamic exchange of partners. *J. Biol. Chem.* **278**, 39265–39268 (2003).
18. Madison, D. L. & Lundblad, J. R. C-terminal binding protein and poly(ADP) ribose polymerase 1 contribute to repression of the p21^{Waf1/Cip1} promoter. *Oncogene* **29**, 6027–6039 (2010).
19. Yeo, C. Q. X. et al. p53 maintains genomic stability by preventing interference between transcription and replication. *Cell Rep.* **15**, 132–146 (2016).
20. Chen, J., Jackson, P. K., Kirschner, M. W. & Dutta, A. Separate domains of p21 involved in the inhibition of Cdk kinase and PCNA. *Nature* **374**, 386–388 (1995).
21. Luo, Y., Hurwitz, J. & Massagué, J. Cell-cycle inhibition by independent CDK and PCNA binding domains in p21^{Cip1}. *Nature* **375**, 159–161 (1995).
22. Mansilla, S. F. et al. Cyclin Kinase-independent role of p21^{CDKN1A} in the promotion of nascent DNA elongation in unstressed cells. *eLife* **5**, e18020 (2016).
23. Uhlen, M. et al. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).
24. Moudry, P. et al. TOPBP1 regulates RAD51 phosphorylation and chromatin loading and determines PARP inhibitor sensitivity. *J. Cell Biol.* **212**, 281–288 (2016).
25. Murai, J. et al. Trapping of PARP1 and PARP2 by clinical PARP inhibitors. *Cancer Res.* **72**, 5588–5599 (2012).
26. Ray Chaudhuri, A. et al. Topoisomerase I poisoning results in PARP-mediated replication fork reversal. *Nat. Struct. Mol. Biol.* **19**, 417–423 (2012).
27. Zellweger, R. et al. Rad51-mediated replication fork reversal is a global response to genotoxic treatments in human cells. *J. Cell Biol.* **208**, 563–579 (2015).
28. Kunkel, T. A. DNA replication fidelity. *J. Biol. Chem.* **279**, 16895–16898 (2004).
29. Flach, J. et al. Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. *Nature* **512**, 198–202 (2014).

Acknowledgements We thank W. Dunphy for the treslin antibody, D. Gomez-Cabello for help with Fig. 2j, the Danish Cancer Society, the Novo Nordisk Foundation, the Danish Council for Independent Research, the Swedish Research Council, the Grant Agency of the Czech Republic (17-14743S), the Czech Ministry of Education, Youth and Sports (NPU LO1304; EATRIS-CZ), and the Danish National Research Foundation (project CARD) for grant support.

Reviewer information *Nature* thanks A. Vindigni and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.M.-M., P.M. and J.B. conceived the study and designed experiments. A.M.-M., P.M., J.M.M.-M., R.S. and M.H.L. performed experiments. A.M.-M., P.M., J.M.M.-M. and J.B. wrote the manuscript. All authors read and accepted the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0261-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0261-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.B. or A.M.-M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture and drug treatment. Human U2-OS, BJ, HeLa, MDA-MB-436 and OVCAR-5 cell lines were grown in DMEM (Gibco) supplemented with 10% FBS (Gibco) and penicillin/streptomycin (Gibco). Cell lines were purchased from ATCC except OVCAR-5 (a gift from T. C. Hamilton). All cell lines were regularly tested for mycoplasma contamination. Cells were treated with different concentrations of following drugs: olaparib (AZD2281; Astra Zeneca), hydroxyurea (Sigma-Aldrich), HDACi (vorinostat, SAHA, MK0683; Selleckchem), ATRi (AZD6738; Astra Zeneca), cisplatin (Hospira), veliparib (ABT-888; Selleckchem), VP-16 (etoposide; Sigma-Aldrich).

DNA fibres. Cell cultures transfected with siRNA and/or treated with different drugs were pulse-labelled with 25 μ M Cl_dU (Sigma-Aldrich) for 20 min, followed by a gentle wash with fresh pre-warmed medium and the second pulse of 250 μ M IdU (Sigma-Aldrich) for 20 min. Labelled cells were collected and DNA fibre spreads prepared as previously described³⁰. For all experimental conditions, five slides were stretched and two or three slides for each condition were stained. Cl_dU was detected first with the rat anti-BrdU antibody (Serotec, OBT0030) and IdU with the mouse anti-BrdU antibody (Becton Dickinson, 347580). Secondary antibodies were DyLight 550 anti-rat (Thermo Fisher Scientific) and Alexa Fluor 488 anti-mouse (Invitrogen), respectively. Images of well-spread DNA fibres were acquired using a LSM700 confocal microscope (Carl Zeiss) and a Plan-Apochromat 63 \times /1.4 numerical aperture (NA) oil immersion objective (Carl Zeiss). Images were acquired semi-automatically by using software autofocus and tile-arrays. Double-labelled replication forks were analysed manually using LSM ZEN software. For each slide, 50–100 forks were scored and fork measurements pooled together for each experiment. At least one more independent experiment was run and, if equivalent experiments were not different statistically, the total number of DNA fibres from both experiments is presented. DMSO-treated control cells or cells transfected with luciferase siRNA as non-targeting control were included for every experiment. Only for experiments presented in Extended Data Fig. 7j, k, the pulse-labelling time was 10 min for Cl_dU and 30 min for IdU. Raw data for every single DNA fibre measurement and statistical analysis are provided in the Source Data.

Fork density. Cells were labelled with a single pulse of 10 μ M BrdU for 20 min and their DNA stretched as described above. DNA was detected using the mouse anti-DNA antibody (Chemicon, MAB 3868), the Alexa Fluor 488 anti-mouse secondary antibody and was additionally stained with Yoyo-1 (Molecular Probes). BrdU was detected using the rat anti-BrdU antibody and the DyLight 550 anti-rat secondary antibody. Well-stretched fibres were used to count the number of forks per Mbp of DNA as previously described³¹. For the origin-to-origin experiment, we stained six different slides for each condition.

Replication program. For single labelling, cells cultured on coverslips were incubated for 30 min in 10 μ M BrdU-containing medium. Cells were then fixed, DNA was denatured for 30 min in 2 M HCl and BrdU was detected using the mouse anti-BrdU antibody and an appropriate secondary antibody. High content microscopy scan^{^R} (Olympus) was used for the S phase analysis. S phase patterns were clustered as early, mid or late from scattering plots BrdU versus DNA content. For double labelling, cells grown on coverslips were pulse-labelled for 30 min with 10 μ M BrdU, washed and incubated in fresh medium for 4 h. After the chase time, cells were incubated for 30 min with 10 μ M EdU. Cells were then fixed and EdU was detected first using the Click-iT EdU Alexa Fluor 568 Imaging Assay kit (Life Technologies) according to the manufacturer's protocol. BrdU was detected as for single-labelled cells. Replication patterns were identified as previously described³⁰ and S phase progression was analysed as previously described³².

BrdU foci detection without denaturation. Cells cultured on coverslips were incubated with 10 μ M BrdU for 48 h to visualize ssDNA. For the last 24 h, cells were treated with 10 μ M olaparib. BrdU incorporated into whole DNA was detected as above, without the HCl denaturation step. The number of BrdU foci in cyclin-A-positive cells was scored using high-throughput microscopy analysis.

Immunofluorescence. Cells cultured on coverslips were fixed with 4% cold formaldehyde (15 min, room temperature), permeabilized with 0.2% Triton X-100 (5 min, room temperature), washed with PBS and incubated with primary antibodies for 1 h at room temperature. Following the washing step, coverslips were incubated with goat anti-rabbit or goat anti-mouse Alexa Fluor 488, 568 or 647 secondary antibodies (Invitrogen) for 1 h at room temperature, washed again with PBS and mounted using Vectashield mounting reagent with DAPI (Vector Laboratories). The primary antibodies used were: γ H2AX (Millipore, 05-636), RAD51 (Abcam, ab63801), cyclin A (Santa Cruz, sc-751; Leica, NCL-CYCLINA), RPA32 (Abcam, ab2175), RPA32 S4/8 (Bethyl, A300-245A), 53BP1 (Santa Cruz, sc-22760), PAR (Trevigen, 4336-BPC), PARP1 (Abcam, ab32138; Trevigen, 4338-MC-50), p21 (Santa Cruz, sc-756 and sc-6246) and p53 (Santa Cruz, sc-126). Images were acquired using a

LSM510 confocal microscope (Carl Zeiss) mounted on an inverted microscope Axiovert 100M (Carl Zeiss), equipped with a Plan-Apochromat 63 \times /1.4 NA oil immersion objective (Carl Zeiss). Image acquisition and analysis were performed using LSM ZEN software. Automated, multi-channel image acquisition was performed using a high-content screening station scan^{^R} (Olympus), equipped with a motorized IX81 microscope (Olympus), an UPlanSApo 40x/0.95 air immersion objective (Olympus) and a digital monochrome C9100 electron multiplying charge coupled device camera (Hamamatsu). Image acquisition and analysis were performed using scan^{^R} acquisition and analysis software (Olympus). Presented results are from 2–4 independent experiments.

EdU detection in G2 cells. Cells grown on coverslips were pulse-labelled for 30 min with 10 μ M EdU. Cells were fixed, permeabilized and washed as above and incubated with the H3 pSer10 (Millipore, 06-570) primary antibody for 1 h at room temperature. After washing with PBS, coverslips were stained with the Alexa Fluor 488 secondary antibody for 1 h (room temperature). EdU was detected using the Click-iT EdU Alexa Fluor 568 Imaging Assay kit (Life Technologies). The Click-iT reaction was carried out for 20 min at room temperature. Coverslips were washed, mounted and viewed under the LSM700 confocal microscope.

Immunoblotting. Whole-cell extracts were prepared in Laemmli sample buffer (50 mM Tris, pH 6.8, 100 mM dithiothreitol (DTT), 2% SDS, 0.1% bromophenol blue and 10% glycerol), separated by SDS-PAGE and transferred to the nitrocellulose membranes (GE Healthcare). The membranes were blocked in 5% dry milk in 0.1% Tween-20 in PBS and probed with primary antibodies. After incubation with horseradish peroxidase (HRP)-conjugated secondary antibodies (Vector Laboratories and Santa Cruz Biotechnology), proteins were visualized using ECL detection reagents (GE Healthcare). The primary antibodies used were: RPA32 (Abcam, ab2175), RPA32 pT21 (Abcam, ab61065), RPA32 S4/8 (Bethyl, A300-245A), RPA32 S33 (Novus, NB100-544), BRCA1 (Santa Cruz, sc-6954), CHK1 (Santa Cruz, sc-8408), CHK1 S317 (Cell Signaling, 2344), CHK2 T68 (Cell Signaling, 2661), Histone H3 (Abcam, ab1791), IMP-B (Abcam, ab2811), PARP1 (Abcam, ab32138), PARP2 (Active Motif, 39743), CHK2 (Santa Cruz, sc-9064), TICRR (from W. G. Dunphy), MTBP (Abcam, ab34704), α -tubulin (Gene Tex, GTX628802), MDM2 (Santa Cruz, sc-813), p21 (Santa Cruz, sc-756 and sc-6246), p53 (Santa Cruz, sc-6243), β -actin (Sigma-Aldrich, A1978) and γ H2AX (Abcam, ab22551).

TUNEL assay. Cells grown on coverslips were pre-extracted with 0.2% Triton X-100 (2 min, room temperature), fixed with 4% cold formaldehyde (15 min, room temperature) and permeabilized with 0.25% Triton X-100 (20 min, room temperature). Cells in Fig. 1g were stained using the In situ BrdU-Red DNA Fragmentation (TUNEL) assay kit (Abcam), following a modified manufacturer's instruction. The TdT reaction was carried out for 1.5 h at 37 °C. BrdU was detected using the mouse anti-BrdU and Alexa Fluor 488 antibodies. Cells were co-stained with the PCNA (Immuno Concepts, 2037) and Alexa Fluor 568 antibodies. Cells in Fig. 1h were labelled using the Click-iT TUNEL Alexa Fluor 488 Imaging Assay kit (Life Technologies) according to the manufacturer's instruction. Images were acquired using the LSM700 confocal microscope and analysed using LSM ZEN software. Presented results are from 2–4 independent experiments.

Comet assay. Cells were collected and resuspended in PBS at the density of 7,500 per μ l. Aliquots of 10 μ l were mixed with 100 μ l of 37 °C LMA (low melting point agarose in PBS, Lonza) and spotted on the slides pre-coated with NMA (normal melting point agarose in ddH₂O, Lonza). Slides covered with coverslips were stored flat for 10 min at 4 °C. Coverslips were gently removed and slides were immersed in cold alkaline or neutral lysis buffer for 2 h at 4 °C. Slides were then washed with cold alkaline or neutral electrophoresis buffer for 20 min at 4 °C and electrophoresis was performed for 25 min at 4 °C (15 V). Slides were then stored flat for 10 min at 4 °C, washed with cold PBS (10 min, 4 °C) and cold ddH₂O (10 min, 4 °C). Slides were dehydrated by successive 5 min washes in cold graded ethanol (70, 96, 99.9%) at 4 °C, air-dried and stored at room temperature. On the following day, slides were rehydrated in ddH₂O and stained with Sybr Gold (1:10,000 in TE buffer; Thermo Fisher Scientific) for 5 min (room temperature). After washing with PBS, slides were mounted with Vectashield mounting reagent (Vector Laboratories). Images were acquired using a fluorescent microscope (Carl Zeiss), a 40 \times air immersion objective (Carl Zeiss) and Comet Assay IV software (Perceptive Instruments). Presented results are from two independent experiments (biological replicates). Alkaline lysis buffer: 1.2 M NaCl, 100 mM Na₂EDTA, 0.1% sodium lauryl sarcosinate, 0.26 M NaOH (pH >13, 4 °C, prepared fresh); alkaline electrophoresis buffer: 0.03 M NaOH, 2 mM Na₂EDTA (pH 12.3, 4 °C). Neutral lysis buffer: 2% sarkosyl, 0.5 M Na₂EDTA, 0.5 mg ml⁻¹ proteinase K (pH 8.0, 4 °C); neutral electrophoresis buffer: 90 mM Tris buffer, 90 mM boric acid, 2 mM Na₂EDTA (pH 8.5, 4 °C)³³.

RNA interference. All siRNA transfections were performed using Lipofectamine RNAiMAX (Invitrogen) according to the manufacturer's instructions. All siRNA duplexes were purchased from Ambion: siBRCA1 (s458), siTICRR #1 (s40362), siTICRR #2 (s40361), siTICRR #3 (s40363), siMTBP #1 (s25786), siMTBP #2 (s25787), siPARP1 (s1097), siPARP2 (S19504), siLIG1

(s8174), siFEN1 (s5194), siTopBP1 (s21823), sip21 (s416), sip53 (Operon; 5'-GACTCCAGTGGTAATCTAC-3'), siLuc as non-targeting (NT) control (MWG; 5'-CGUACGCGGAAUACUUCGA-3'). Mission shRNA plasmids were purchased from Sigma-Aldrich. Lentiviruses carrying shRNA were generated and purified using standard techniques. To obtain U2-OS cells stably transduced with *p21* (*CDKN1A*) shRNA (TRCN287021 and TRCN287091) or control shRNA (SHC016), cells were selected by puromycin (1 µg ml⁻¹) treatment for 5 days before experiments were carried out.

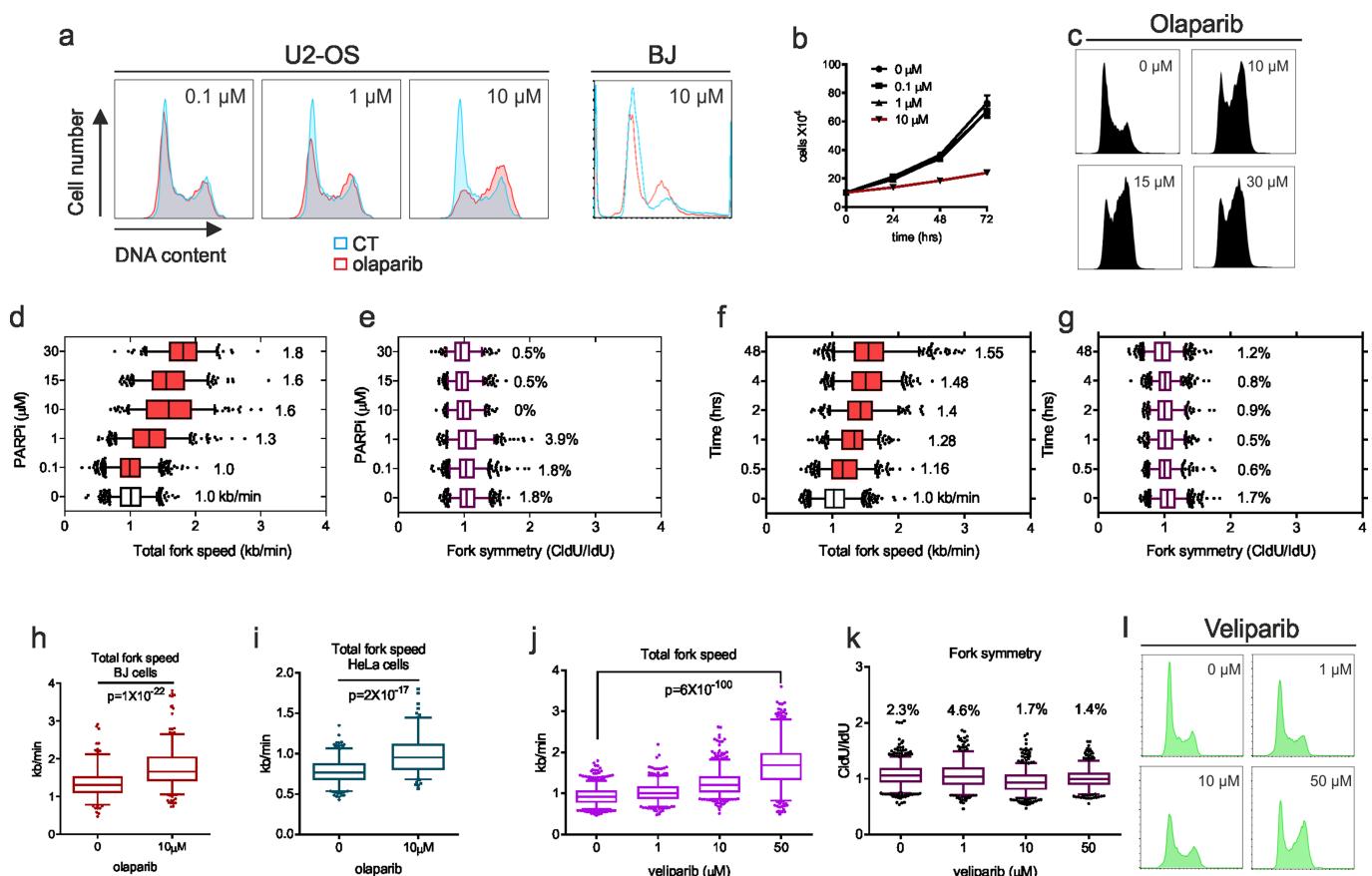
MNase assay. Cells (5 × 10⁶) were washed with cold PBS and scraped in 5 ml of NBA buffer (85 mM KCl, 5.5% sucrose, 10 mM Tris, pH 7.5, 0.2 mM EDTA, 0.5 mM spermidine, 250 µM PMSF and protease inhibitors) with 0.5% Igepal CA-630 (Sigma-Aldrich). Cells were collected, mixed with 5 ml of NBA buffer and incubated on ice for 5 min. Nuclei were recovered by centrifugation at 350g for 4 min and resuspended in 0.5 ml of NBC buffer (NBA without EDTA). Nuclei were digested with 1,000 U of micrococcal nuclease in 1 × MNase buffer and aliquots were taken at 0, 1, 2.5, 5 and 7 min, respectively. DNA was purified with proteinase K and phenol:chloroform. DNA was quantified and equal amounts were resolved in 1.2% agarose gel.

Cell cycle analysis. Cells transfected with siRNA and/or treated as described above were fixed with 70% ethanol at -20 °C and incubated on ice for at least 30 min. Afterwards, cells were washed with cold PBS and labelled for 5 min (RT) with 10 µg ml⁻¹ of propidium iodide (Invitrogen) in PBS, containing 5 µg ml⁻¹ RNase A (Life Technologies). Cells were analysed immediately on FACSVerse (Becton Dickinson) and acquired data were analysed using the Cell Cycle platform of FlowJo software.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

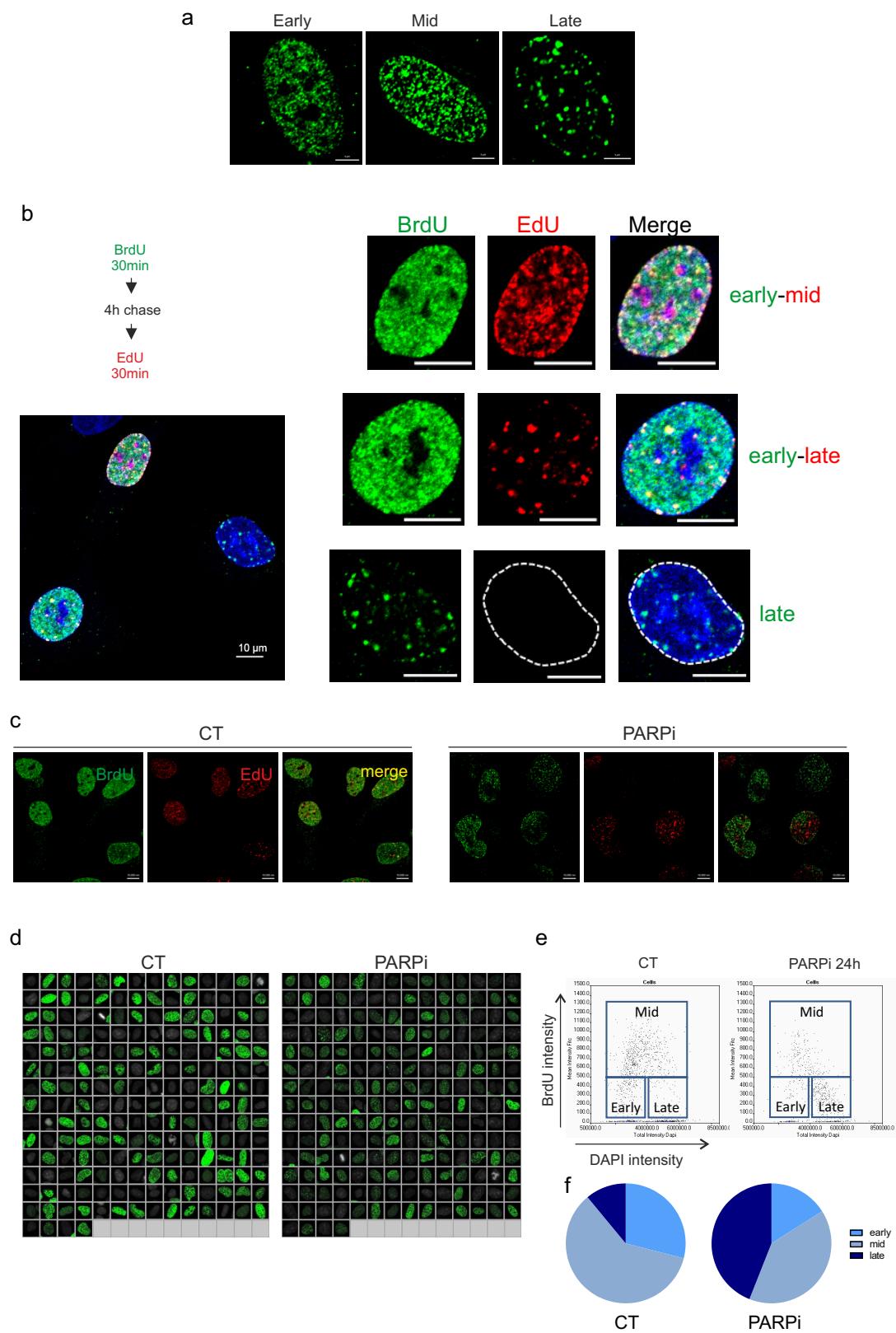
Data availability. Data are available from the corresponding author(s) upon reasonable request.

30. Maya-Mendoza, A., Olivares-Chauvet, P., Kohlmeier, F. & Jackson, D. A. Visualising chromosomal replication sites and replicons in mammalian cells. *Methods* **57**, 140–148 (2012).
31. Sheu, Y. J., Kinney, J. B., Lengronne, A., Pasero, P. & Stillman, B. Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression. *Proc. Natl Acad. Sci. USA* **111**, E1899–E1908 (2014).
32. Kohlmeier, F., Maya-Mendoza, A. & Jackson, D. A. EdU induces DNA damage response and cell death in mESC in culture. *Chromosome Res.* **21**, 87–100 (2013).
33. Olive, P. L. & Banáth, J. P. The comet assay: a method to measure DNA damage in individual cells. *Nat. Protocols* **1**, 23–29 (2006).
34. Pines, A., Mullenders, L. H., van Attikum, H. & Luijsterburg, M. S. Touching base with PARPs: moonlighting in the repair of UV lesions and double-strand breaks. *Trends Biochem. Sci.* **38**, 321–330 (2013).
35. Sims, J. L., Berger, S. J. & Berger, N. A. Poly(ADP-ribose) polymerase inhibitors preserve nicotinamide adenine dinucleotide and adenosine 5'-triphosphate pools in DNA-damaged cells: mechanism of stimulation of unscheduled DNA synthesis. *Biochemistry* **22**, 5188–5194 (1983).
36. Dutto, I. et al. *p21*^{CDKN1A} regulates the binding of poly(ADP-ribose) polymerase-1 to DNA repair intermediates. *PLoS ONE* **11**, e0146031 (2016).
37. Breslin, C. et al. The XRCC1 phosphate-binding pocket binds poly (ADP-ribose) and is required for XRCC1 function. *Nucleic Acids Res.* **43**, 6934–6944 (2015).
38. Ray Chaudhuri, A., Ahuja, A. K., Herrador, R. & Lopes, M. Poly(ADP-ribosyl) glycohydrolase prevents the accumulation of unusual replication structures during unperturbed S phase. *Mol. Cell. Biol.* **35**, 856–865 (2015).
39. Strzalka, W. & Ziemiennowicz, A. Proliferating cell nuclear antigen (PCNA): a key factor in DNA replication and cell cycle regulation. *Ann. Bot.* **107**, 1127–1140 (2011).



Extended Data Fig. 1 | Fork acceleration is PARPi dose- and time-dependent, and cell-type independent. **a**, Cell cycle profiles of U2-OS cells treated with different concentrations of the PARPi olaparib (0.1, 1 or 10 μ M) and BJ cells treated with 10 μ M olaparib for 24 h; $n = 3$ biological replicates. **b**, Number of U2-OS cells treated with different concentrations of olaparib (0.1, 1 or 10 μ M) for 72 h; $n = 3$ biological replicates. Drug was refreshed every 24 h. Data are mean \pm s.d. **c**, Cell cycle profiles of U2-OS cells treated with increasing concentrations of the olaparib (10, 15 or 30 μ M) for 24 h; $n = 2$ biological replicates. **d**, DNA fibres from U2-OS cells 24 h after treatment with increasing concentrations of olaparib (0.1, 1, 10, 15 or 30 μ M). Scored forks: 0 μ M PARPi = 503; 0.1 μ M = 606; 1 μ M = 406; 10 μ M = 244; 15 μ M = 372; 30 μ M = 217; $n = 2$ biological replicates. Mean fork speed (kb min^{-1}) is indicated next to each condition. **e**, ClDU/IdU ratios calculated from values in **d**. Percentage of highly asymmetric forks (ClDU/IdU ratios < 0.5 and > 1.5) is indicated next to each condition. **f**, DNA fibres from U2-OS cells treated with 10 μ M olaparib for different periods of time (0.5, 1, 2, 4 or 48 h). Scored forks:

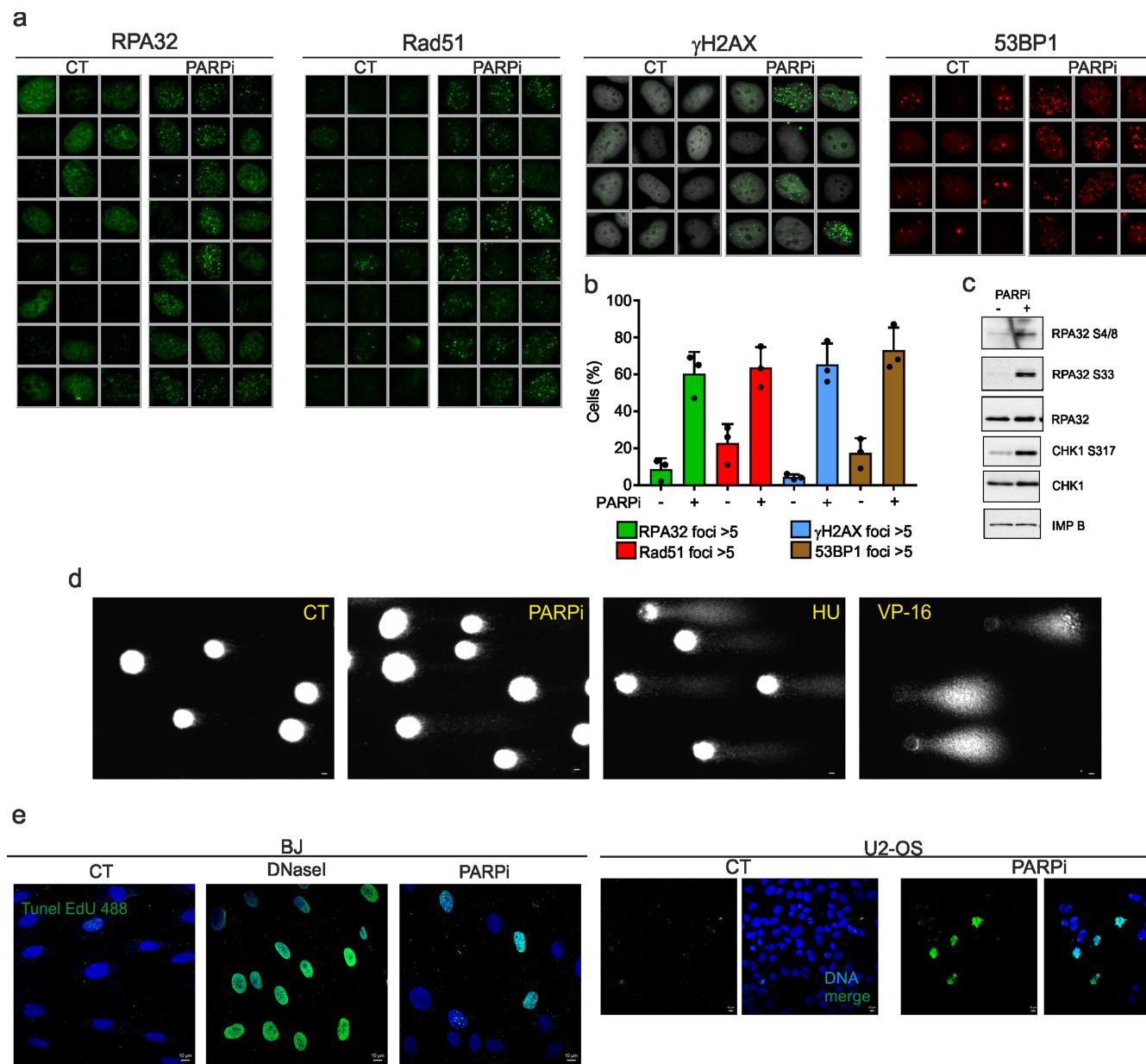
0 h = 744; 0.5 h = 450; 1 h = 379; 2 h = 314; 4 h = 465; 48 h = 589; $n = 2$ biological replicates. Mean fork speed is indicated. **g**, ClDU/IdU ratios calculated from values in **f**. Percentage of highly asymmetric fork is indicated. **h**, DNA fibres from BJ cells treated with 10 μ M olaparib for 24 h. Scored forks: 0 μ M = 198; 10 μ M = 317; $n = 2$ biological replicates. **i**, DNA fibres from HeLa cells treated with 10 μ M olaparib for 4 h. Scored forks: 0 μ M = 285; 10 μ M = 142; $n = 2$ technical replicates. **j**, DNA fibres from U2-OS cells 24 h after treatment with increasing concentrations of veliparib (1, 10 or 50 μ M). Scored forks: 0 μ M = 689; 1 μ M = 408; 10 μ M = 571; 50 μ M = 408; $n = 2$ biological replicates. **k**, ClDU/IdU ratios calculated from values in **j**. Percentage of highly asymmetric forks is indicated above each condition. **l**, Cell cycle profiles of U2-OS cells treated with increasing concentrations of veliparib (1, 10 or 50 μ M) for 24 h; $n = 2$ biological replicates. For box plots in **d–k**, whiskers indicate the fifth and ninetieth percentiles, and the centre values depict the median.

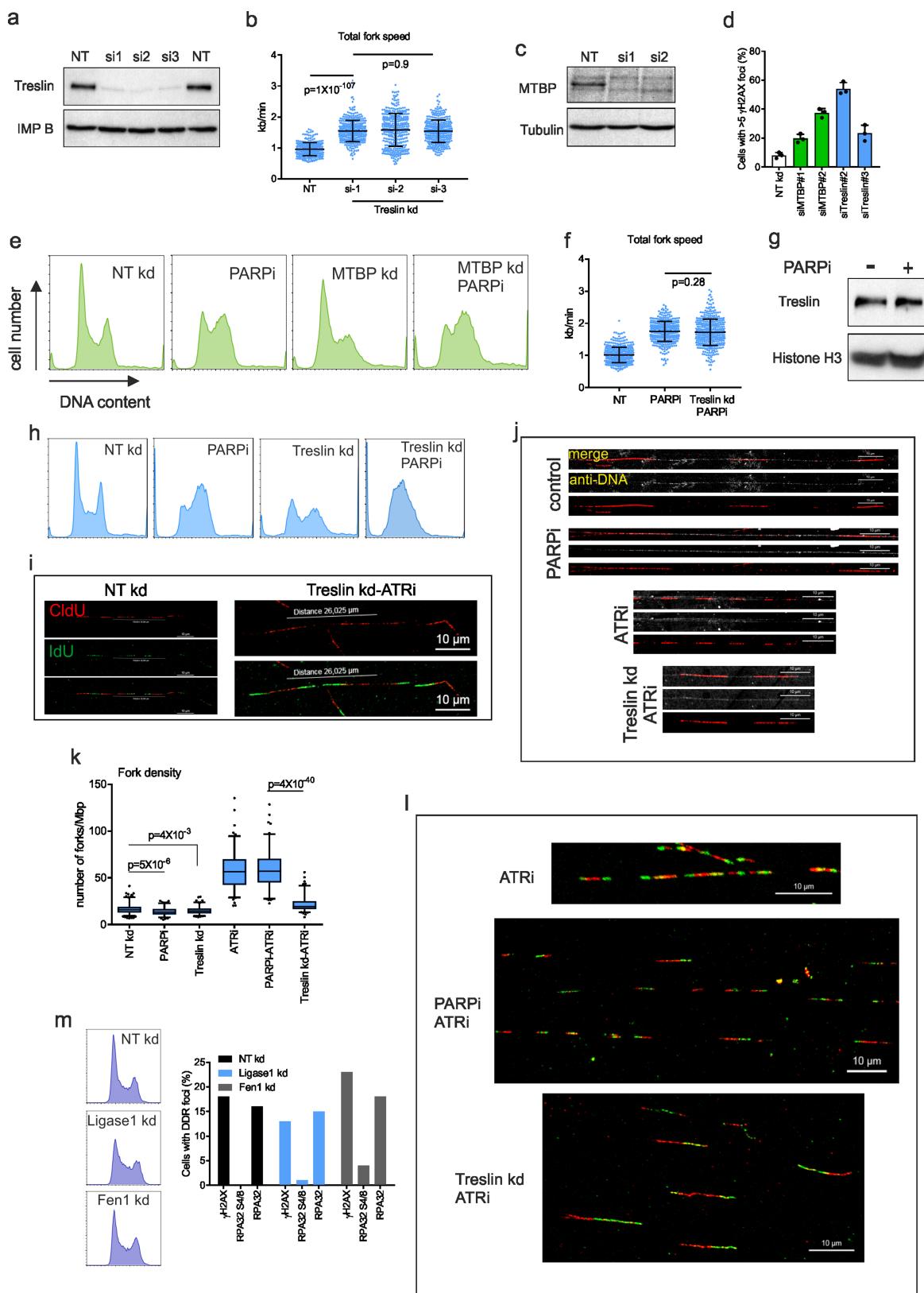


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | PARPi induces accumulation of cells in mid-late S phase. **a**, Representative images of DNA replication patterns in control U2-OS cells. Cells were pulse-labelled with 10 μ M BrdU for 30 min. Scale bars, 5 μ m. **b**, Outline of the experimental design of detailed DNA replication pattern analysis. U2-OS cells were labelled with BrdU (green) for 30 min, washed, chased for 4 h in fresh medium and labelled with EdU (red) for 30 min. Transition between replication patterns was classified as early–early (cells that did not leave early S phase during the experiment time), early–mid and mid–late (cells that progressed to the consecutive part of S phase) and early–late (cells that progressed fast through S phase). Scale bars, 10 μ m. **c**, Representative images of double-labelled DNA

replication patterns in U2-OS cells treated with DMSO (CT) or 10 μ M olaparib (PARPi) for 24 h. Scale bars, 10 μ m. **d**, Representative images of BrdU-positive nuclei from control and PARPi-treated U2-OS cells in early, mid and late S phase. Images were acquired using high-throughput microscopy. **e**, Percentage of U2-OS cells in early, mid and late S phase after treatment with DMSO or 10 μ M olaparib for 24 h was quantified using high-throughput microscopy based on BrdU intensity versus DNA content. S phase patterns were gated as indicated. **f**, Distribution of S phase patterns in U2-OS cells treated with DMSO or 10 μ M olaparib for 24 h (CT: early = 29%; mid = 60%, late = 11%; PARPi: early = 16%; mid = 40%, late = 44%; $n = 2$ biological replicates) (see Source Data).

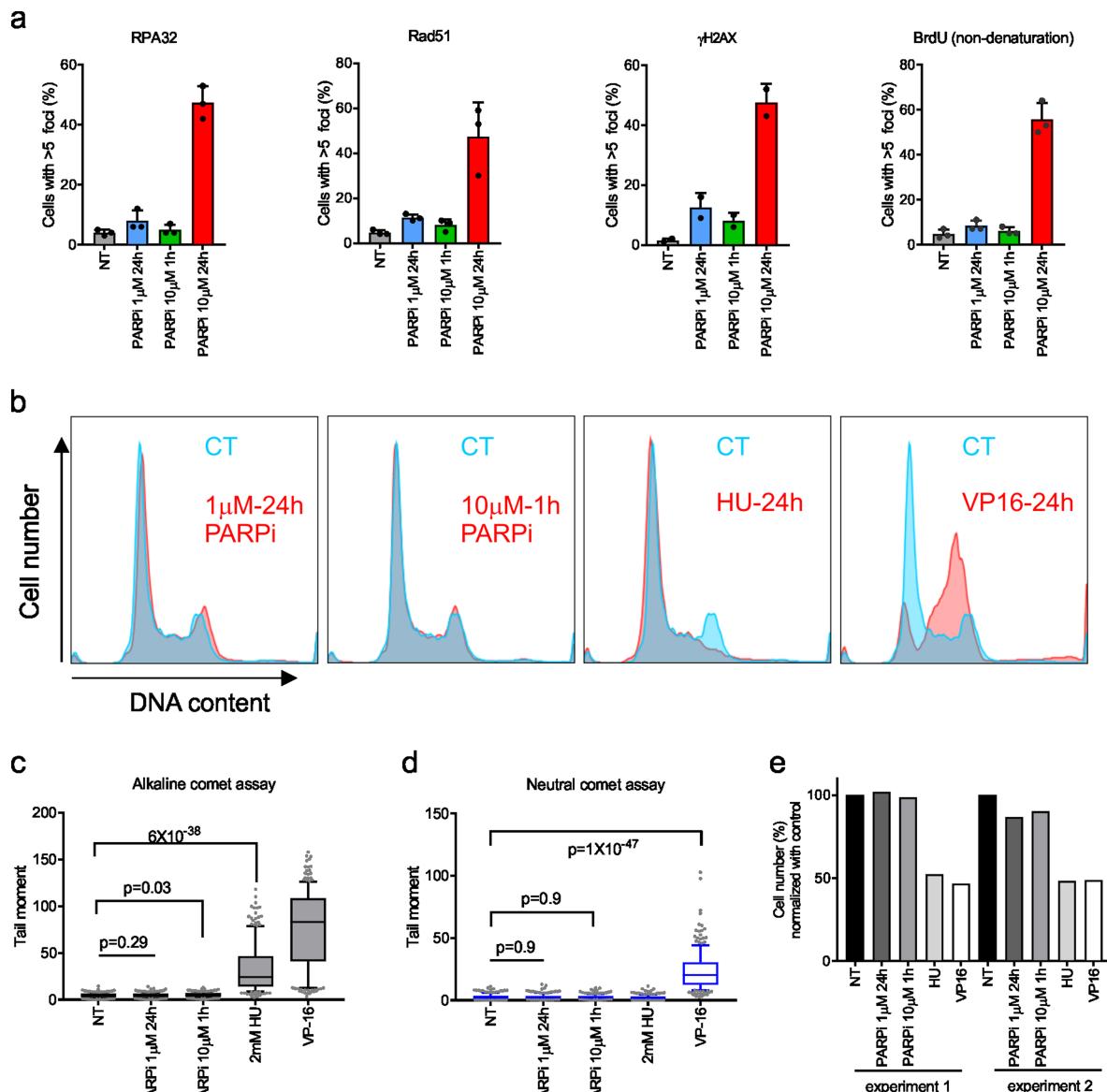




Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Fork speed and origin activation. **a**, Immunoblots of treslin knockdown efficiency in U2-OS cells, 72 h after transfection with three different siRNAs (si1–si3). $n = 2$ biological replicates. **b**, DNA fibres from non-targeting or treslin-knockdown U2-OS cells, 72 h after transfection with three different siRNA. Mean fork speed (kb min^{-1}): NT = 0.96; si1 = 1.54; si2 = 1.58; si3 = 1.54. Scored forks: NT = 316; si1 = 367; si2 = 368; si3 = 358; $n = 2$ biological replicates. Data are mean \pm s.d. P values determined by Welch's two-tailed t -test (see Source Data). **c**, Immunoblots of MTBP knockdown efficiency in U2-OS cells, 72 h after transfection with two different siRNAs. Tubulin is a loading control; $n = 2$ biological replicates. **d**, Percentage of non-targeting, MTBP- or treslin-knockdown U2-OS cells with more than γ H2AX foci. Data are mean \pm s.d., $n = 2$ biological replicates. **e**, Cell cycle profiles of non-targeting or MTBP-knockdown U2-OS cells. Indicated cells were treated with 10 μM PARPi for 24 h; $n = 2$ biological replicates. **f**, DNA fibres from non-targeting or treslin-knockdown U2-OS cells. Indicated cells were treated as in **e**. Mean fork speed (kb min^{-1}): NT = 1.0; PARPi = 1.74; treslin KD–PARPi = 1.72. Scored forks: NT = 410; PARPi = 395; treslin KD–PARPi = 424; $n = 2$ biological replicates. Data are mean \pm s.d. P values determined by Welch's two-tailed t -test. **g**, Immunoblots of the chromatin-associated fraction of treslin after 24 h of treatment with 10 μM olaparib. Histone H3 is a loading control; $n = 2$ biological

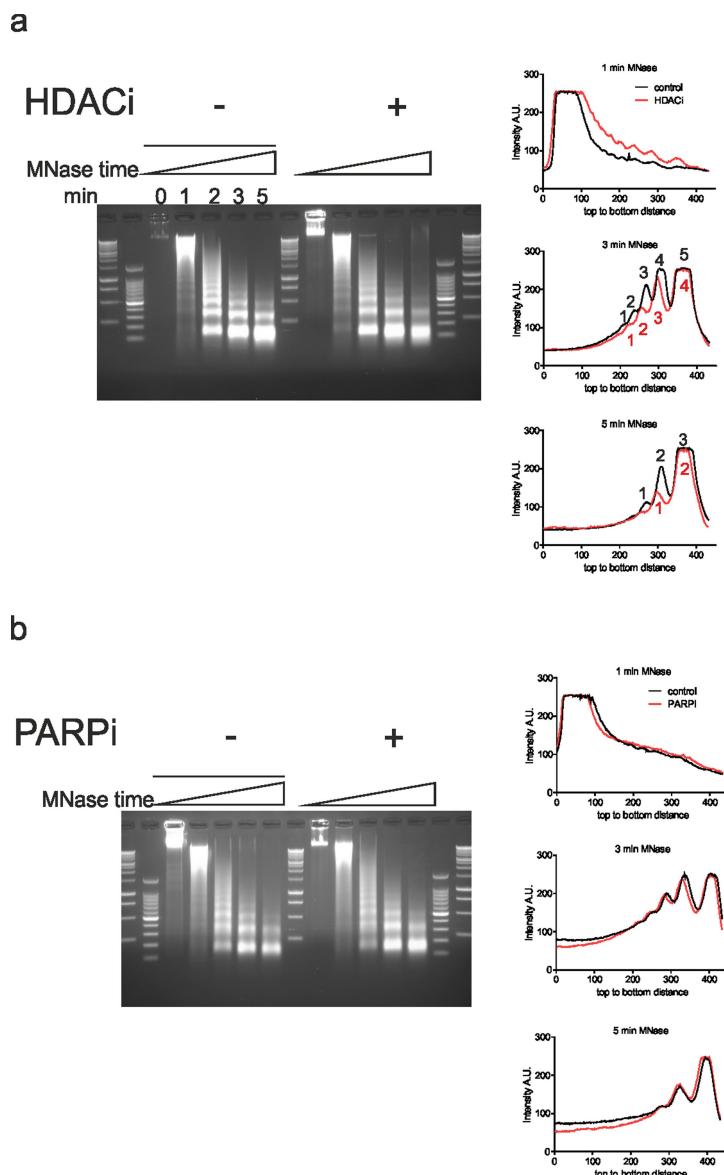
replicates. **h**, Cell cycle profiles of non-targeting or treslin-knockdown U2-OS cells. Indicated cells were treated as in **e**; $n = 2$ biological replicates. **i**, Representative images of the origin-to-origin distance measurements in double-labelled DNA fibres from non-targeting or treslin-knockdown/ATR-inhibited U2-OS cells. Scale bars, 10 μm . **j**, Representative images of fork density from non-targeting or treslin-knockdown U2-OS cells. Indicated cells were treated with 10 μM PARPi for 24 h or 1 μM ATRi for 1 h before a 20-min 10 μM BrdU pulse. DNA (grey) and BrdU (red) were detected by immunofluorescence. Scale bars, 10 μm . **k**, Number of forks in a well-spread single DNA fibre, counted from multiple fibres per each condition and converted into number of forks per Mb: NT = 17; PARPi = 14; treslin-KD = 15; ATRi = 58; PARPi–ATRi = 59; treslin KD–ATRi = 22; $n = 3$ biological replicates. Whiskers indicate fifth and ninety-fifth percentiles and centre values depict the median. P values determined by Welch's two-tailed t -test. **l**, Representative images of double-labelled DNA fibres from non-targeting or treslin-knockdown U2-OS cells. Indicated cells were treated with 10 μM PARPi for 24 h and/or 1 μM ATRi for 1 h, before pulse-labelling with CldU (red) for 20 min and IdU (green) for another 20 min. Scale bars, 10 μm . **m**, Left, cell cycle profiles of non-targeting, LIG1- or FEN1-knockdown U2-OS cells. Right, the percentage of U2-OS cells with more than five DDR foci (representative experiment from $n = 2$ biological replicates).



Extended Data Fig. 5 | Low-dosage of olaparib did not induce strong DDR. **a**, Percentage of U2-OS cells with more than 5 DDR foci after treatment with 1 μ M or 10 μ M olaparib for 24 h, or 10 μ M olaparib for 1 h. Data are mean \pm s.d., n = 3 biological replicates (see Source Data).

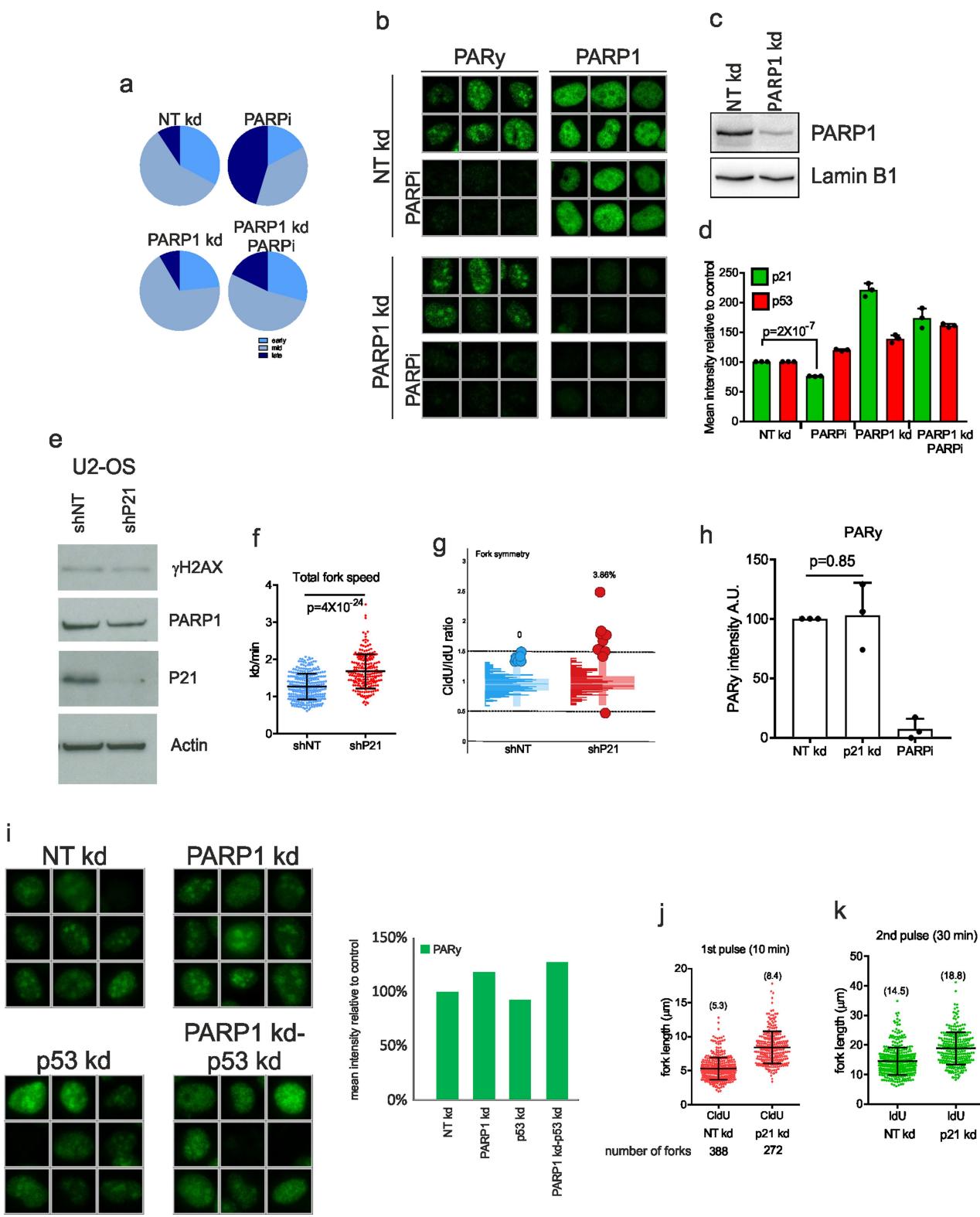
b, Cell cycle profiles of U2-OS cells treated with 1 μ M olaparib for 24 h, or 10 μ M olaparib for 1 h. HU (2 mM, 24 h) was included as a positive control for inhibition of S phase progression, and VP-16 (10 μ M, 24 h) for G2/M

phase arrest (n = 2 biological replicates). **c, d**, Alkaline (c) or neutral (d) comet assays from U2-OS cells treated as in **b**. HU is a positive control for ssDNA; VP-16 is a positive control for dsDNA. Whiskers indicate fifth and ninety-fifth percentiles, and centre value depicts the median. P values determined by two-sided Kolmogorov–Smirnov test and two-tailed t -test; n = 2 biological replicates. **e**, Number of U2-OS cells treated as in **b** relative to control cells (n = 2 biological replicates).



Extended Data Fig. 6 | Olparib did not induce global changes in chromatin structure. **a**, Analysis of chromatin sensitivity to MNase digestion in control and HDAC-inhibited cells ($n = 3$ biological replicates; representative experiment is shown). Gel densitometries at different time points are presented next to the agarose gel. The number of detected bands

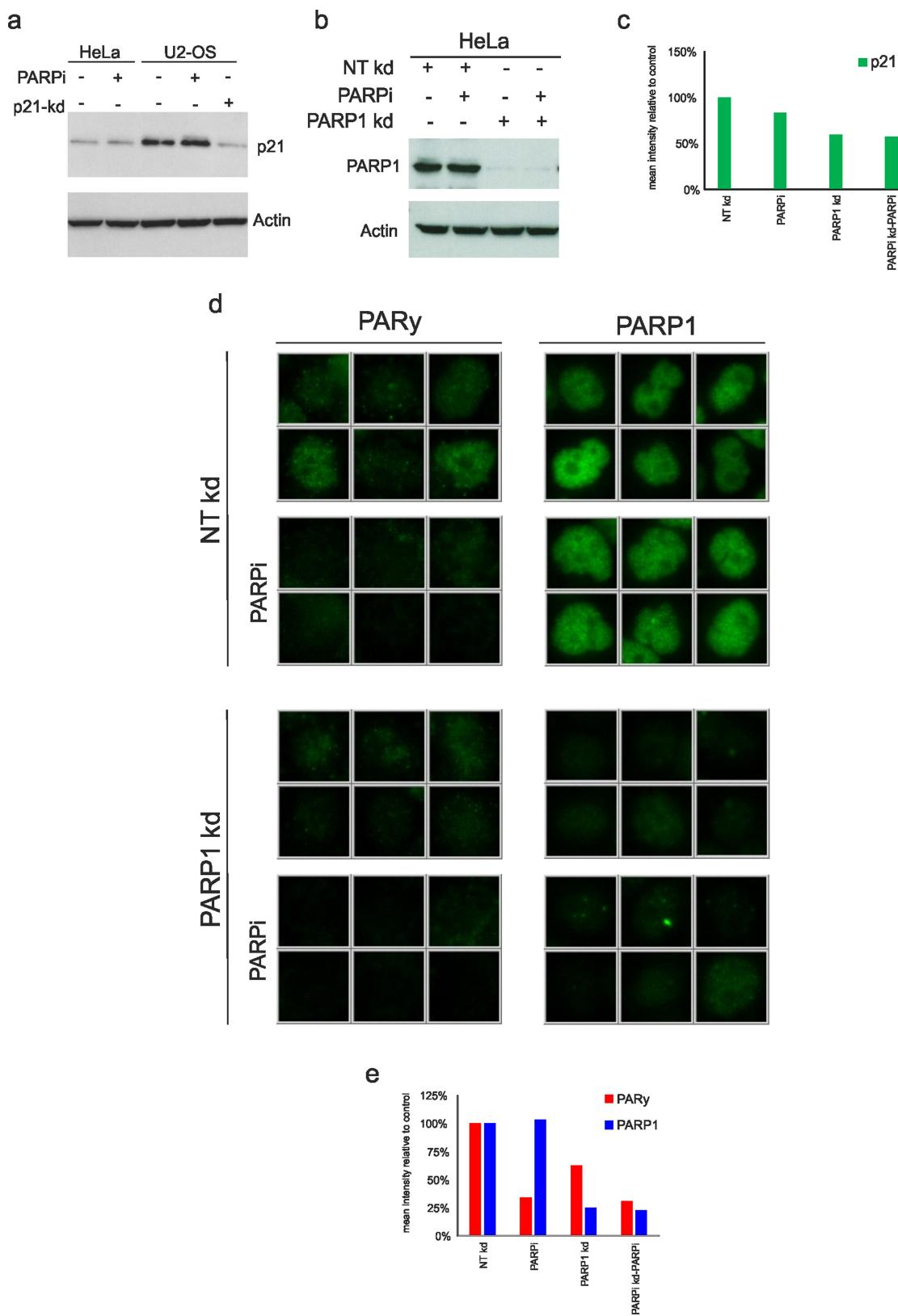
is shown on densitometry plots. The smaller number of bands, the more sensitive chromatin is. **b**, Analysis of chromatin sensitivity to MNase digestion in control and PARP-inhibited cells ($n = 3$ biological replicates; representative experiment is shown).



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | p21 and fork speed regulation. **a**, Distribution of S phase patterns by BrdU incorporation in non-targeting or PARP1-knockdown U2-OS cells. Indicated cells were treated with 10 μ M PARPi for 24 h; $n = 3$ biological replicates (see Source Data). **b**, Representative images of PAR and PARP1 in non-targeting or PARP1-knockdown U2-OS cells. Indicated cells were treated as in **a**. **c**, Immunoblots of PARP1 knockdown efficiency in U2-OS cells 72 h after transfection with siRNA. Lamin B1 is a loading control; $n = 2$ biological replicates. **d**, Mean intensity of p21 and p53 in non-targeting or PARP1-knockdown U2-OS cells. Indicated cells were treated as in **a**. Data are mean \pm s.d. P values were determined by a two-tailed t -test; $n = 3$ biological replicates. **e**, Immunoblots of γ H2AX, PARP1 and p21 in the p21-knockdown U2-OS stable cell line. Actin is a loading control; $n = 2$ biological replicates. shNT, U2-OS cell line with non-targeting shRNA. **f**, DNA fibres from the p21-knockdown U2-OS stable cell line. Mean fork speed (kb min^{-1}): shNT = 1.2; shP21 = 1.68. Scored forks: shNT = 305; shP21 = 207; $n = 2$

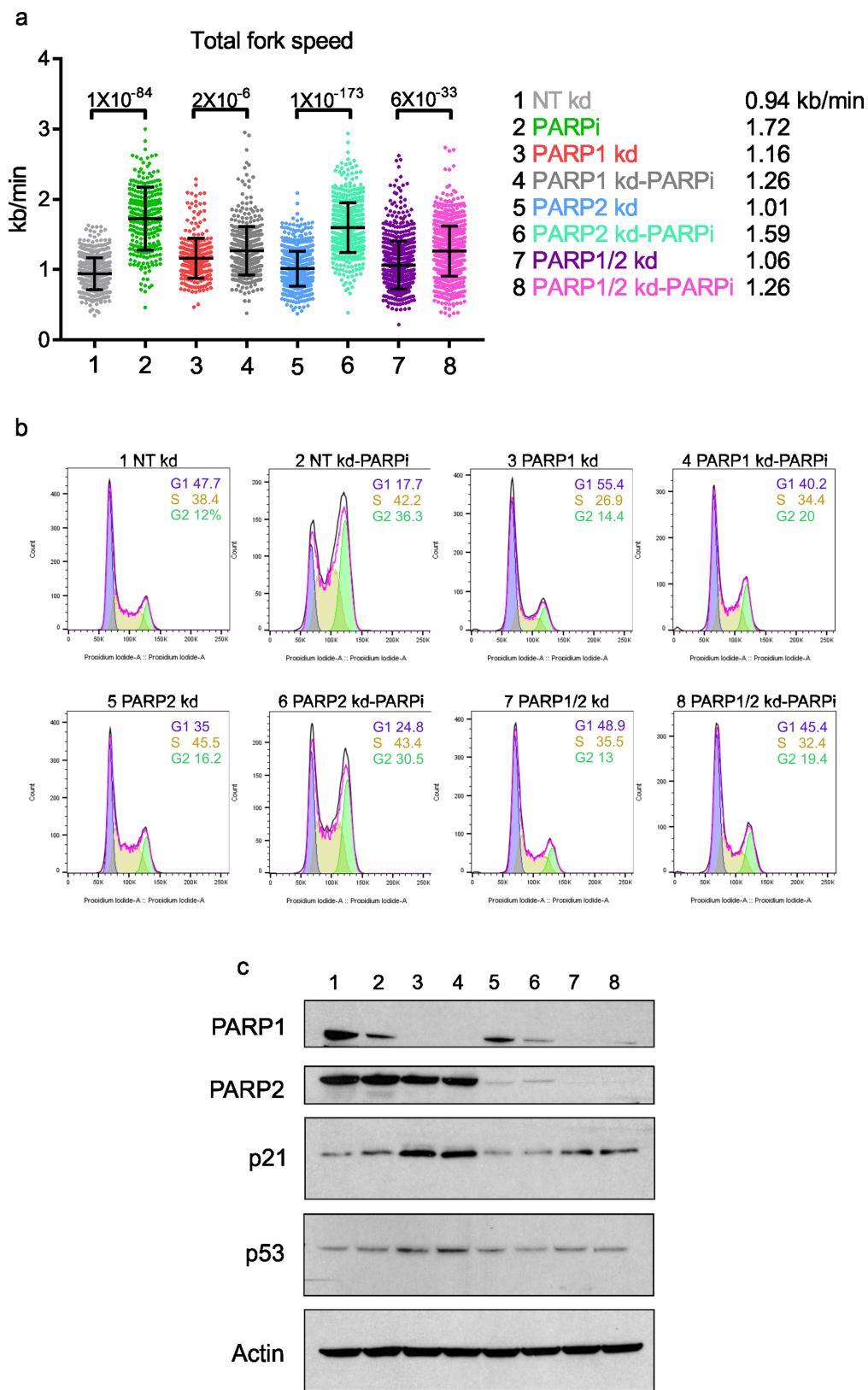
biological replicates. Data are mean \pm s.d. P values determined by two-tailed Welch's t -test. **g**, CldU/IdU ratios calculated from values in **f**. Percentage of highly asymmetric forks (CldU/IdU ratios < 0.5 and > 1.5) is indicated above each condition. **h**, Mean intensity of PAR in non-targeting or p21-knockdown U2-OS cells. Indicated cells were treated as in **a**. Data are mean \pm s.d. P values determined by two-tailed Welch's t -test; $n = 3$ biological replicates. **i**, Representative images of PAR in non-targeting, PARP1- or p53-knockdown U2-OS cells. Mean intensity of PAR relative to non-targeting control in U2-OS cells (representative results from $n = 2$ biological replicates). **j**, U2-OS cells, 72 h after transfection with non-targeting or *p21* siRNA were pulse-labelled for 10 min with CldU (red), washed and pulse-labelled with IdU (green) for 30 min. Fork length (μm) of the first (CldU) pulse. **k**, Fork length (μm) of the second (IdU) pulse from the experiment in **j**. Mean \pm s.d. of separate forks is indicated above each condition. Scored forks: NT = 388; p21 KD = 272; $n = 2$ biological replicates.



Extended Data Fig. 8 | Effect of PARP1 knockdown in HeLa cells.

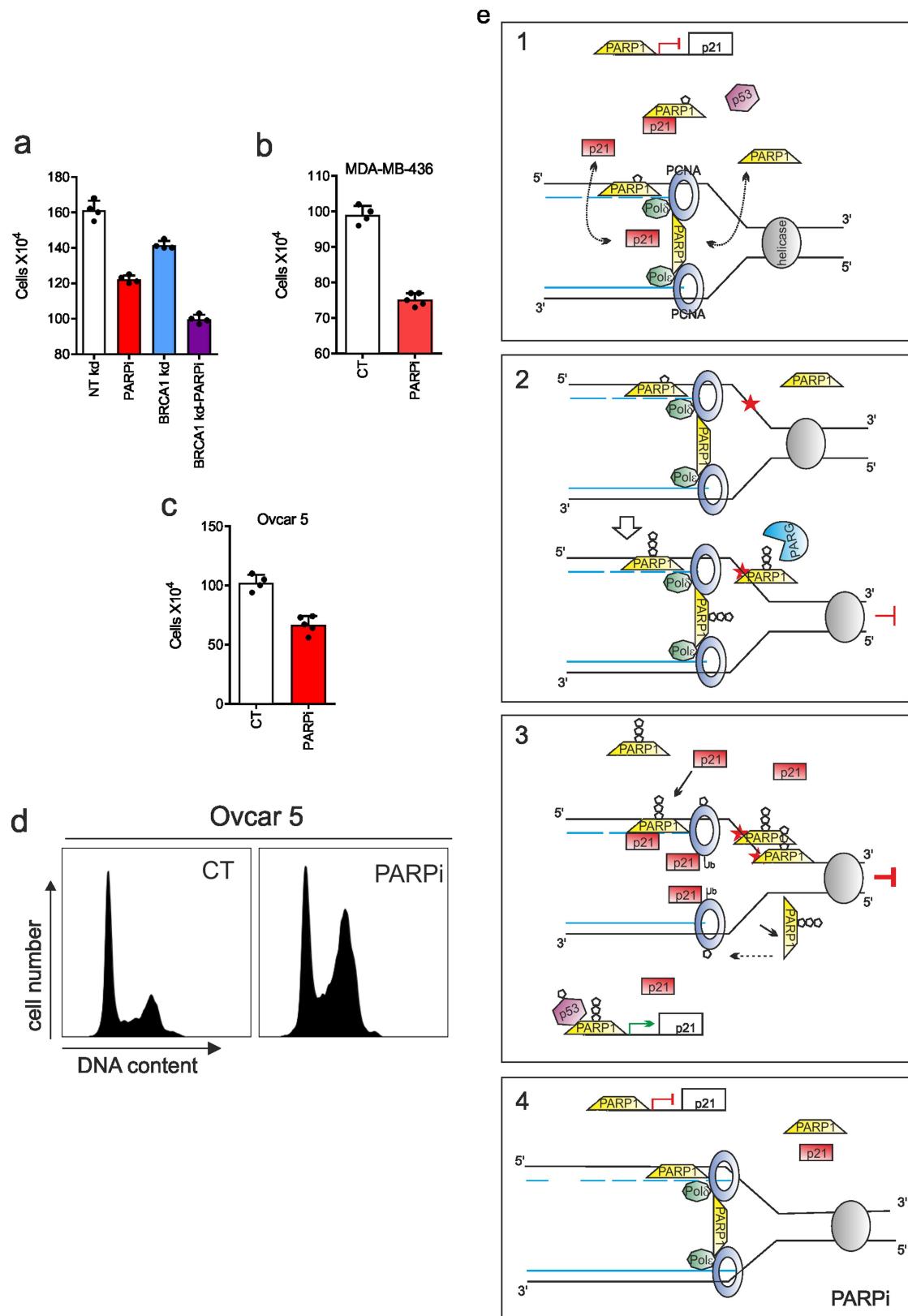
a, Immunoblots of p21 in HeLa cells and in non-targeting or p21-knockdown U2-OS cells (lines 2, 4; 10 μ M PARPi, 24 h). $n = 2$ biological replicates. **b**, Immunoblots of PARP1 in non-targeting or PARP1-knockdown HeLa cells (lines 2, 4; 10 μ M PARPi, 24 h). $n = 2$ biological replicates. **c**, Mean intensity of p21 in non-targeting or PARP1-knockdown HeLa cells. Indicated cells were treated with PARPi (10 μ M, 24h),

representative experiment from $n = 2$ biological replicates (see Source Data). **d**, Representative images of PAR and PARP1 in non-targeting or PARP1-knockdown HeLa cells. Indicated cells were treated as in **c**. **e**, Mean intensity of PAR and PARP1 in non-targeting or PARP1-knockdown HeLa cells. Indicated cells were treated as in **c**; representative experiment from $n = 2$ biological replicates.



Extended Data Fig. 9 | Fork speed in double-knockdown PARP1/2. **a**, DNA fibres from U2-OS cells 72 h after transfection with different siRNAs. Indicated cells were treated with 10 μ M PARPi for 24 h. Mean fork speed (kb min^{-1}) is indicated. Scored forks: NT = 586; PARPi = 263; PARP1 KD = 327; PARP1 KD-PARPi = 451; PARP2 KD = 794; PARP2 KD-PARPi = 597; PARP1/2 KD = 831; PARP1/2 KD-PARPi = 962; $n = 2$ biological replicates. Data are mean \pm s.d. P values were determined

by two-tailed Welch's t -test (see Source Data). **b**, Cell cycle profiles of U2-OS cells 72 h after transfection with different siRNA and treated as in **a**. Percentage of cells in different phases of the cell cycle analysed using FlowJo software are indicated next to the histograms; $n = 2$ biological replicates. **c**, Immunoblots of PARP1, PARP2, p21 and p53 from experimental conditions described in **a**. $n = 2$ biological replicates.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | The FSRN. **a**, Number of non-targeting or BRCA1-knockdown U2-OS cells. Indicated cells were treated with 10 μ M PARPi for 24 h. Data are mean \pm s.d. $n = 4$ biological replicates (see Source Data). **b**, Number of MDA-MB-436 BRCA1-deficient cells 24 h after olaparib treatment. Data are mean \pm s.d., $n = 4$ biological replicates. **c**, Number of OVCAR-5 ovarian cancer cells 24 h after olaparib treatment. Data are mean \pm s.d., $n = 4$ biological replicates. **d**, Cell cycle profiles of OVCAR-5 ovarian cancer cells 24 h after olaparib treatment. **e**, The fork speed regulatory network (FSRN) model. (1) During unperturbed S phase, inactive PARP1 inhibits transcription of *p21*. Induction of PARP enzymatic activity is necessary for *p21* promoter activation, by relief of repression, for both p53-dependent and -independent pathways (our data and shown previously¹⁸). PARP1 has high affinity to DNA nicks and ssDNA. Binding of PARP1 to DNA nicks stimulates its activity³⁴. Moreover, a steady-state level of PARylation is necessary for the normal cell physiology, as excess of PARP activity after DNA damage reduces the amount of NAD⁺, affecting the ATP level³⁵. PARP1 can bind directly to *p21* and the PARP inhibitor olaparib reduces this interaction³⁶. In our model, levels of p53 (pink hexagon), *p21* (red rectangle), *p21*-PARP1 complex, free PARP1 (yellow trapezoid) and a low level of PARylation

(small empty pentagon) are maintained at a steady state during normal S phase. PCNA (blue circle) is associated with replication forks and is bound by polymerase (Pol) δ on the lagging strand and Pol ϵ on the leading strand. In replication factories, PARP1 can be associated directly to DNA (that is, at the nicks of the lagging DNA strand) and to PCNA¹⁷. The balance between these players enables the normal speed of fork progression to be maintained. (2) Any break in DNA is promptly recognized by PARP1, which triggers its activity. PARylation can promote the recruitment of important DDR proteins³⁷ or can directly inhibit fork progression. Excess of PARylation needs to be removed by PARG enzymes, allowing the fork to resume³⁸. (3) When DNA is severely damaged, PARP1 is strongly activated. PARylated PARP1 releases *p21* from the *p21*-PARP1 complexes. PARylated PARP1 is also bound by p53, which transactivates *p21*. After prolonged fork arrest, processive DNA polymerases dissociate from modified PCNA³⁹ and are replaced by *p21*. *p21* can inhibit PCNA-dependent DNA replication in the absence of cyclin/CDK. Furthermore, *p21* blocks the ability of PCNA to activate DNA Pol δ ¹⁵. Therefore, PARylation and *p21* act as suppressors of DNA replication. (4) PARP inhibitors disrupt FSRN.

CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions

Michal Zimmermann^{1,14}, Olga Murina^{2,14}, Martin A. M. Reijns², Angelo Agathangelou³, Rachel Challis², Žygimantė Tarnauskaitė², Morwenna Muir⁴, Adeline Fluteau², Michael Aregger⁵, Andrea McEwan¹, Wei Yuan⁶, Matthew Clarke⁶, Maryou B. Lambros⁶, Shankara Panesha⁷, Paul Moss⁸, Megha Chandrashekhar^{5,9}, Stéphane Angers¹⁰, Jason Moffat^{5,9,11}, Valerie G. Brunton⁴, Traver Hart¹², Johann de Bono^{6,13}, Tatjana Stankovic³, Andrew P. Jackson^{2*} & Daniel Durocher^{1,9*}

The observation that BRCA1- and BRCA2-deficient cells are sensitive to inhibitors of poly(ADP-ribose) polymerase (PARP) has spurred the development of cancer therapies that use these inhibitors to target deficiencies in homologous recombination¹. The cytotoxicity of PARP inhibitors depends on PARP trapping, the formation of non-covalent protein–DNA adducts composed of inhibited PARP1 bound to DNA lesions of unclear origins^{1–4}. To address the nature of such lesions and the cellular consequences of PARP trapping, we undertook three CRISPR (clustered regularly interspersed palindromic repeats) screens to identify genes and pathways that mediate cellular resistance to olaparib, a clinically approved PARP inhibitor¹. Here we present a high-confidence set of 73 genes, which when mutated cause increased sensitivity to PARP inhibitors. In addition to an expected enrichment for genes related to homologous recombination, we discovered that mutations in all three genes encoding ribonuclease H2 sensitized cells to PARP inhibition. We establish that the underlying cause of the PARP-inhibitor hypersensitivity of cells deficient in ribonuclease H2 is impaired ribonucleotide excision repair⁵. Embedded ribonucleotides, which are abundant in the genome of cells deficient in ribonucleotide excision repair, are substrates for cleavage by topoisomerase 1, resulting in PARP-trapping lesions that impede DNA replication and endanger genome integrity. We conclude that genomic ribonucleotides are a hitherto unappreciated source of PARP-trapping DNA lesions, and that the frequent deletion of RNASEH2B in metastatic prostate cancer and chronic lymphocytic leukaemia could provide an opportunity to exploit these findings therapeutically.

We carried out dropout CRISPR screens with olaparib in three human cell lines of diverse origins, representing both neoplastic and non-transformed cell types (Fig. 1a, Extended Data Fig. 1a, b). The cell lines selected were HeLa, which is derived from a human papilloma virus-induced cervical adenocarcinoma; RPE1-hTERT, a telomerase-immortalized retinal pigment epithelium cell line; and SUM149PT, originating from a triple-negative breast cancer with a hemizygous *BRCA1* mutation⁶. SUM149PT cells express a partially defective *BRCA1* protein (*BRCA1*-Δ11q)⁷ and thus provided a sensitized background to search for enhancers of PARP-inhibition cytotoxicity in cells that have compromised homologous recombination. The screens were performed in technical triplicates, and a normalized depletion score for each gene was computed using DrugZ⁸. To identify high-confidence hits, we used a stringent false discovery rate (FDR) threshold of 1% in one cell line. To this initial list, we added genes that

were found at an FDR threshold of less than 10% in at least two cell lines. This analysis identified 64, 61 and 116 genes, the inactivation of which caused sensitization to olaparib in the HeLa, RPE1-hTERT and SUM149PT cell lines, respectively, giving a total of 155 different genes (Supplementary Table 1).

Out of this list, 13 genes scored positive in all three cell lines and a further 60 genes were common to two cell lines, which we combine to define a core set of 73 high-confidence PARP inhibitor (PARPi)-resistance genes (Fig. 1b, Supplementary Table 1). Gene Ontology analysis of the 73- and 155-gene sets (Fig. 1c, Extended Data Fig. 1c, respectively) shows strong enrichment for biological processes related to homologous recombination, providing unbiased confirmation that the screens identified bona fide regulators of the response to PARP inhibition. Mapping the 73-gene set on the HumanMine protein–protein interaction data (Fig. 1d) generated a highly connected network consisting of DNA damage response genes that include many regulators of homologous recombination (such as *BRCA1*, *BARD1*, *BRCA2* and *PALB2*), components of the Fanconi anaemia pathway, as well as the kinases *ATM* and *ATR*. Outside or at the edge of the network, we noted the presence of genes encoding the *MUS81*–*EME1* nuclease, splicing and general transcription factors (such as *SF3B1/5* and *CTDP1*) and the three genes coding for the ribonuclease (RNase) H2 enzyme complex (*RNASEH2A*, *RNASEH2B* and *RNASEH2C*). *RNASEH2A*, *RNASEH2B* and *RNASEH2C* were hits in all three cell lines, with *RNASEH2A* and *RNASEH2B* being the two highest-scoring genes, as determined by the mean DrugZ value from the three cell lines (Supplementary Table 1). A similar analysis of the 155-gene set generated an even denser network, with additional genes lying at the periphery of a homologous recombination and Fanconi anaemia core (Extended Data Fig. 1d).

Next, we generated RNase H2-null HeLa, RPE1, SUM149PT and HCT116 clonal cell lines using genome editing (denoted as KO; Extended Data Fig. 2a–d) and confirmed that RNase H2 deficiency caused hypersensitivity to both olaparib and a second clinical-stage PARPi, talazoparib, in all cell lines tested (Fig. 2a, b, Extended Data Fig. 2e–g, with half-maximal effective concentration (EC_{50}) values reported in Extended Data Fig. 2h). The *RNASEH2A*^{KO} and *RNASEH2B*^{KO} cells also exhibited increased levels of apoptosis after PARP inhibition (Extended Data Fig. 2i–l), a phenotype that was particularly prominent with talazoparib treatment (Extended Data Fig. 2i–l). Given the strength of the PARPi-induced phenotypes in RNase H2-deficient cells, and as RNase H2 had not previously been linked to the response to PARP inhibition, we sought to determine the mechanism of PARPi sensitization in RNase H2-deficient cells.

¹The Lunenfeld–Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. ²MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. ³Institute for Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK. ⁴Cancer Research UK Edinburgh Centre, University of Edinburgh, Edinburgh, UK. ⁵Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. ⁶The Institute of Cancer Research, London, UK. ⁷Heartlands Hospital, Birmingham, UK. ⁸Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK. ⁹Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. ¹⁰Department of Pharmaceutical Sciences, Leslie Dan Faculty of Pharmacy & Department of Biochemistry, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada. ¹¹Canadian Institute for Advanced Research, Toronto, Ontario, Canada. ¹²Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ¹³Royal Marsden NHS Foundation Trust, London, UK. ¹⁴These authors contributed equally: Michal Zimmermann, Olga Murina. *e-mail: andrew.jackson@igmm.ed.ac.uk; durocher@lunenfeld.ca

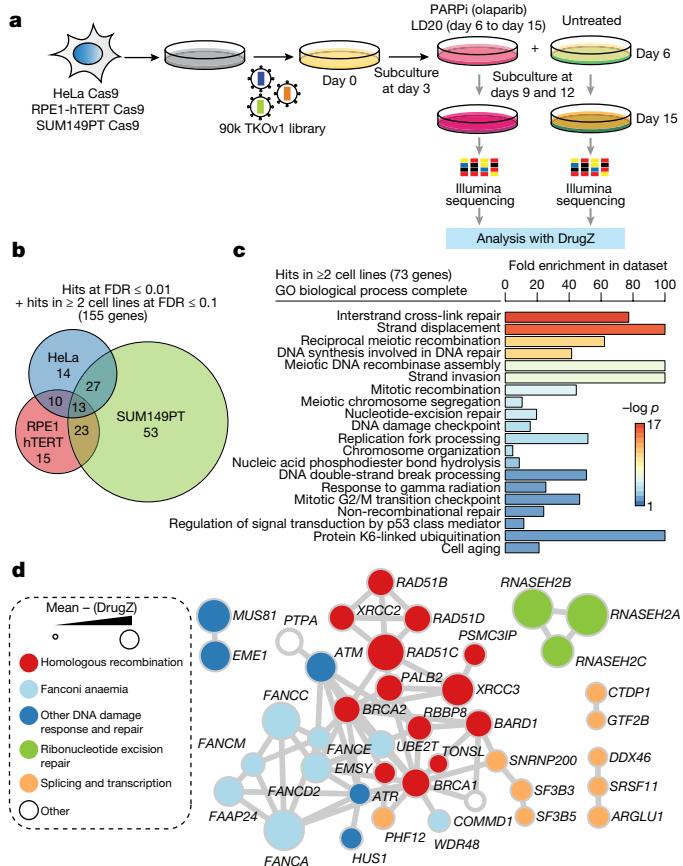


Fig. 1 | CRISPR screens identify determinants of PARPi sensitivity. **a**, Schematic of the screening pipeline. **b**, Venn diagram of all high-confidence hits ($FDR \leq 0.01$ in one cell line and $FDR \leq 0.1$ in at least two cell lines) in individual cell lines. **c**, Gene Ontology (GO) terms significantly ($P < 0.05$, binomial test with Bonferroni correction) enriched among hits common to at least two cell lines. **d**, esyN network analysis of interactions between hits common to at least two cell lines. Node size represents the mean DrugZ score across cell lines. Out of 73 genes, 31 are mapped on the network. See also Extended Data Fig. 1.

As deficiency in homologous recombination causes PARPi sensitivity, we first considered that RNase H2 might promote homologous recombination. Consistent with this possibility, fission yeast cells that combine mutations in RNase H2 and RNase H1 have defects in homologous recombination⁹. However, in RNase H2-deficient cells, RAD51 readily formed ionizing radiation-induced foci, suggesting efficient recombinase filament assembly (Fig. 2c, d, Extended Data Fig. 3a, b). Furthermore, the efficiency of homologous recombination, as assessed by the direct repeat-green fluorescent protein (DR-GFP) assay¹⁰, was at near wild-type levels in cells transduced with *RNASEH2A* and *RNASEH2B* short guide RNAs (sgRNAs) (Fig. 2e, Extended Data Fig. 3c, d). Third, rather than presenting reduced homologous recombination, *RNASEH2A*^{KO} cells displayed higher levels of sister chromatid exchanges, reminiscent of the 'hyper-rec' phenotype observed in RNase H2-deficient yeast¹¹ (Fig. 2f). This phenotype was probably caused by increased levels of replication-dependent DNA damage, as determined by γ -H2AX staining (Fig. 2g, Extended Data Fig. 3e–h) and marked poly(ADP-ribosylation) of PARP1 (Fig. 2h, Extended Data Fig. 3i, j), supporting previous observations of replication-associated genome instability in yeast and mammalian cells deficient in RNase H2^{12–14}.

The increased levels of sister chromatid exchanges prompted us to test whether RNase H2-deficient cells required homologous recombination for survival. Indeed, we observed synthetic lethality when an sgRNA against *RNASEH2B* was delivered into engineered *BRCA1*^{KO} and *BRCA2*^{KO} cell lines in the RPE1-hTERT and DLD-1 backgrounds, respectively (Fig. 2i, Extended Data Fig. 3k–o).

RNase H2 cleaves single ribonucleotides incorporated into DNA, as well as longer RNA–DNA hybrids¹⁵. To distinguish between these two functions, we carried out cellular complementation experiments with variants of RNase H2. The sensitivity of *RNASEH2A*^{KO} cells to olaparib was not rescued by either a catalytically-inactive RNase H2 enzyme (*RNASEH2A*(D34A/D169A)), or by a separation-of-function mutant (*RNASEH2A*(P40D/Y210A)¹⁶) that retains activity against RNA–DNA hybrids, but not DNA-embedded monoribonucleotides (Fig. 2j, Extended Data Fig. 4). These data indicate that it is probably the removal of genome-embedded ribonucleotides by ribonucleotide excision repair (RER), and not RNA–DNA hybrid cleavage by RNase H2, that protects cells from PARPi-induced cytotoxicity.

To determine the genetic basis of the sensitivity of *RNASEH2A*^{KO} cells to PARPi, we carried out CRISPR screens to identify mutations that restored resistance to talazoparib in RNase H2-deficient HeLa and RPE1-hTERT cell lines (Fig. 3a, Extended Data Fig. 5a, Supplementary Table 2). The screens identified a single common gene, *PARP1*. The genetic dependency on *PARP1* for talazoparib- and olaparib-induced cytotoxicity was confirmed in double mutant *RNASEH2A*^{KO} *PARP1*^{KO} cells (Fig. 3b, Extended Data Fig. 5b–e), providing evidence that the lethality associated with PARP inhibition requires formation of trapped PARP1–DNA adducts⁴. Consistent with this finding, treatment with veliparib, a PARP inhibitor with poor trapping ability⁴ induced much less apoptosis than olaparib or talazoparib in *RNASEH2A*^{KO} cells (Extended Data Fig. 5f).

Analysis of DNA content by flow cytometry revealed that *RNASEH2A*^{KO} cells arrest in S phase in a PARP1-dependent manner upon talazoparib treatment (Fig. 3c, Extended Data Fig. 5g). *RNASEH2A*^{KO} cells also demonstrated increased levels of talazoparib-induced γ -H2AX and these levels did not decline upon drug removal (Fig. 3c, Extended Data Fig. 5h). These observations suggest that unresolved DNA lesions induced by PARP trapping are the likely cause of cell death in PARPi-treated *RNASEH2A*^{KO} cells.

Genome instability in RER-deficient yeast cells is dependent on an alternative, topoisomerase 1 (TOP1)-mediated ribonucleotide excision pathway^{17–19}. In this process, TOP1 enzymatic cleavage 3' of the embedded ribonucleotide results in DNA lesions predicted to engage PARP1, including nicks with difficult-to-ligate 2'-3'-cyclic phosphate ends^{17,18,20} and covalent TOP1–DNA adducts (TOP1 cleavage complexes²¹) in conjunction with single-strand DNA gaps or double-strand breaks²². Given that the mechanisms promoting genome instability in mammalian RNase H2-deficient cells remain poorly defined, we assessed whether TOP1 action on misincorporated ribonucleotides contributed to the DNA damage observed in human RER-deficient cells. Short-term depletion of TOP1 using short interfering RNAs (siRNAs) reduced the number of γ -H2AX foci in RNase H2-deficient cells to nearly wild-type levels (Fig. 3d–f, Extended Data Fig. 6a). Furthermore, TOP1-mediated ribonucleotide cleavage contributed to PARPi sensitivity, as depletion of TOP1 with independent siRNAs in *RNASEH2A*^{KO} cells reduced the levels of talazoparib-induced apoptosis (Fig. 3g, Extended Data Fig. 6b–e). TOP1 depletion also reduced talazoparib-induced apoptosis in the RER-deficient *RNASEH2A*(P40D/Y210A) cells (Extended Data Fig. 6f–h) and ameliorated the talazoparib-induced S-phase arrest (Extended Data Fig. 6i). Together, these results strongly suggest that the processing of genome-embedded ribonucleotides by TOP1 leads to DNA lesions that engage PARP1, creating a vulnerability to PARP trapping.

The *RNASEH2B* gene resides on chromosome 13q14 in proximity to two tumour suppressor loci. One of them, the *DLEU2-mir-15-16* microRNA cluster, is a target of 13q14 deletions observed in over 50% of chronic lymphocytic leukaemia (CLL) cases²³. As a result, collateral homozygous deletion of *RNASEH2B* can occur in CLL and other haematopoietic malignancies²⁴. Additionally, in prostate cancer, frequent deletions at 13q14 involving the *RB1* but not the *BRCA2* locus²⁵ might also result in *RNASEH2B* loss. Such 13q14 deletions are late events associated with endocrine-therapy resistance, luminal-to-basal phenotype transition and rapid disease progression^{26,27}.

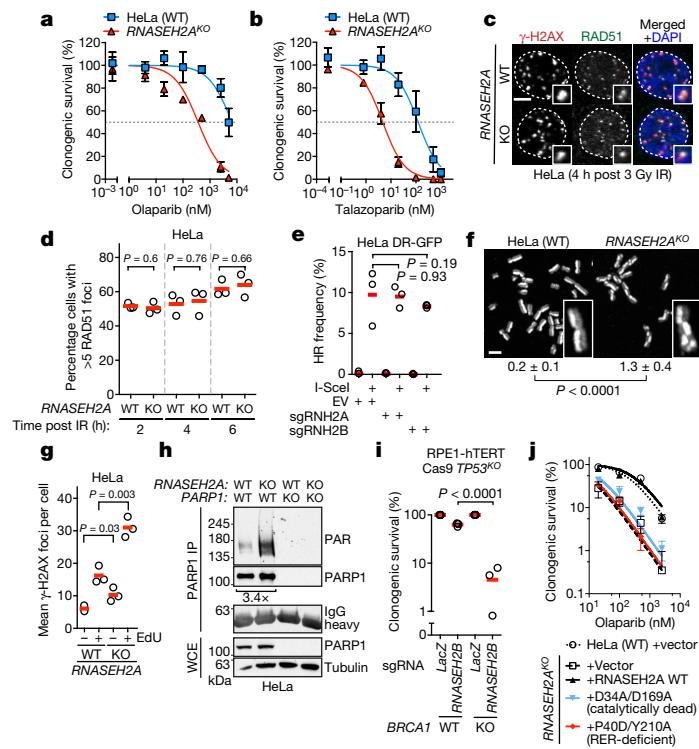


Fig. 2 | Defective RER causes PARPi sensitivity, DNA damage and synthetic lethality with BRCA1 deficiency. **a, b**, Reduced survival of HeLa *RNASEH2A*^{KO} cells after treatment with the indicated PARPi. Data are mean \pm s.d. normalized to untreated cells. Solid lines show a nonlinear least-squares fit to a three-parameter dose–response model. **c–f**, *RNASEH2A*^{KO} cells are homologous-recombination proficient. **c, d**, Normal RAD51 focus formation in *RNASEH2A*^{KO} HeLa cells after ionising radiation (IR) exposure. **c**, Representative micrographs of wild-type (WT) and *RNASEH2A*^{KO} HeLa cells stained with the indicated antibodies. **d**, Quantification of the percentage of cells with more than five RAD51/γ-H2AX colocalizing foci at the indicated time points. Data are from three biologically independent experiments. Scale bar, 10 μ m. **e**, Homologous recombination (HR) is not impaired in RNase H2-null cells. Quantification of gene conversion in DR-GFP reporter cells¹¹ transduced with Cas9 and sgRNAs targeting *RNASEH2A* and *RNASEH2B* (sgRNH2A and sgRNH2B, respectively) or empty vector (EV) with or without I-SceI transfection. Values are normalized to the transfection efficiency of a control GFP vector. **f**, Increased sister chromatid exchanges (SCEs) in *RNASEH2A*^{KO} cells. Representative micrographs of SCEs in wild-type and *RNASEH2A*^{KO} metaphases. Numbers below the images indicate the numbers of SCEs per chromosome. Data are mean \pm s.d. from three biologically independent experiments. Scale bars, 10 μ m. **g, h**, Spontaneous replication-associated damage and increased PARP1 activation in *RNASEH2A*^{KO} cells. **g**, Quantification of mean γ-H2AX

immunofluorescent foci number per nucleus in 5-ethynyl-2'-deoxyuridine (EdU)⁺ and EdU⁻ wild-type and *RNASEH2A*^{KO} cells. **h**, Representative poly(ADP-ribose) (PAR) immunoblot of PARP1 immunoprecipitates (IP) from whole cell extracts (WCEs). Mean fold-increase in poly(ADP-ribosylation) between wild-type and *RNASEH2A*^{KO} indicated. Data are from three biologically independent experiments and are normalized to immunoprecipitated PARP1 levels. Tubulin and IgG heavy chain were included as loading controls. **i**, Synthetic lethality in the combined absence of RNase H2 and BRCA1. Quantification of colony formation of wild-type *BRCA1*-proficient and *BRCA1*^{KO} RPE1-hTERT Cas9 *TP53*^{KO} cells transduced with constructs encoding sgRNAs targeting *LacZ* or *RNASEH2B*. Open circles, individual values normalized to the sgRNA *LacZ* values; red lines, mean. Data are from three biologically independent experiments. **j**, PARPi sensitivity is associated with RER deficiency. Survival of olaparib-treated wild-type and *RNASEH2A*^{KO} HeLa cells transduced with indicated Flag-tagged constructs. Data are mean \pm s.d. from three biologically independent experiments and are normalized to untreated cells. Solid lines, nonlinear least-squares fit to a three-parameter dose–response model. **d, e, g**, Open circles, individual values; red lines, mean from three biologically independent experiments. At least 100 (**d, g**) or 1,000 (**e**) cells were analysed per sample per experiment. *P* values in **d–g** and **i** are from unpaired two-tailed *t*-tests. See also Extended Data Figs. 2–4.

We determined *RNASEH2B* copy number by multiplex ligation-dependent probe amplification (MLPA) in 100 patients with CLL. *RNASEH2B* deletions were present in 43% of CLL samples, with biallelic loss detected in 14%. Co-deletion of the *DLEU2* microRNA cluster was confirmed by comparative genomic hybridization (CGH) microarray (Fig. 4a, Extended Data Fig. 7a, b), establishing that collateral *RNASEH2B* loss is frequent in CLL. Furthermore, analysis of whole-exome sequencing of metastatic castration-resistant prostate cancers (CRPCs)²⁸ demonstrated frequent collateral loss of *RNASEH2B* with *RB1* gene deletion co-occurring in 34% of tumours (2% biallelic loss; Extended Data Fig. 7c).

The frequent collateral deletion of *RNASEH2B* prompted us to test whether *RNASEH2B* loss in cancer cells could be an actionable vulnerability to PARP inhibition. To do so, we performed ex vivo analysis on primary CLL cells derived from 21 of the 100 patient samples assayed above. Patient characteristics of selected samples were similar across groups (Extended Data Table 1). RNase H2 status was confirmed by

enzymatic assay of CLL lysates (Fig. 4b) and short-term CLL cultures were established from peripheral blood leukocyte samples by stimulating their proliferation with IL21 and co-culture with CD40-ligand expressing MEFs (Extended Data Fig. 8a, b, Supplementary Fig. 2). *RNASEH2B*-deficient cells were found to be significantly more sensitive to PARPi and especially to talazoparib, with the degree of sensitivity correlating with number of *RNASEH2B* alleles lost (Fig. 4c, Extended Data Fig. 8c).

We then asked whether RNase H2 deficiency also confers PARPi sensitivity to tumours in xenograft experiments, using isogenic HCT116 cells with and without *RNASEH2A* deletion (Extended Data Fig. 2a, g, l). Cells were implanted in the flanks of CD-1 nude mice and, following establishment of tumours, mice were treated with talazoparib given its higher trapping activity. While talazoparib treatment did not lead to tumour regression, we observed significantly higher sensitivity to talazoparib in tumours lacking RNase H2 (Fig. 4d). Furthermore, a second xenograft experiment confirmed this sensitivity to be specific

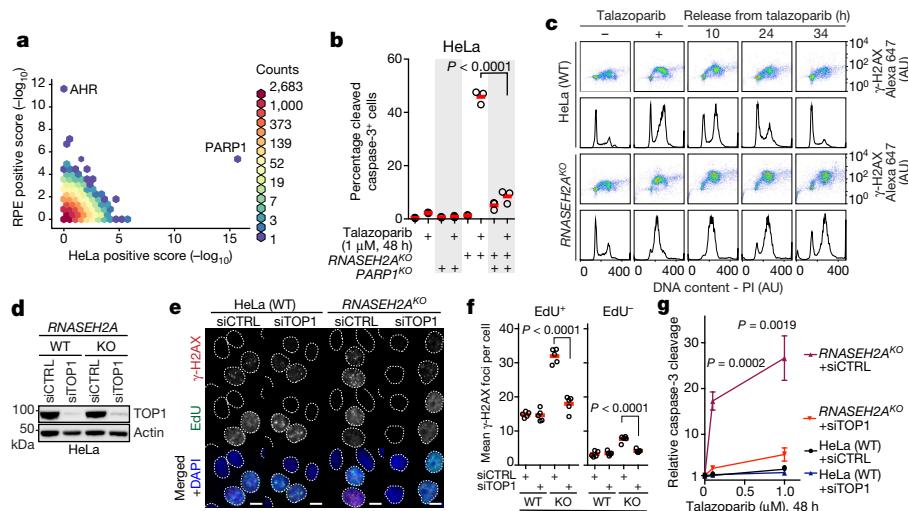


Fig. 3 | PARPi-induced PARP1 trapping occurs in RER-deficient cells as a result of TOP1-mediated processing of genomic ribonucleotides. **a, b**, PARP1 is required for PARPi-induced toxicity in *RNASEH2A*^{KO} cells. **a**, CRISPR screens for talazoparib sensitivity suppressors in *RNASEH2A*^{KO} HeLa Cas9 and RPE1 Cas9 *TP53*^{KO} cell lines. Model-based analysis of genome-wide CRISPR–Cas9 knockout (MAGECK) positive scores for each gene plotted. Colours indicate gene density in each hexagonal bin. **b**, Percentage of cleaved caspase-3⁺ cells of the indicated genotype with or without talazoparib treatment as measured by flow cytometry (FACS). Open circles, individual experiments; red lines, mean from three biologically independent experiments. **c**, DNA damage persists on withdrawal of PARPi in *RNASEH2A*^{KO} cells. Wild-type and *RNASEH2A*^{KO} HeLa cells were treated with talazoparib and released into fresh medium for the indicated times before being processed for γ-H2AX immunofluorescence and propidium iodide (PI) staining. The γ-H2AX immunofluorescence (pseudocolor plots) and cell cycle (histograms) FACS profiles shown are representative of three biologically independent experiments. **d–f**, Increased γ-H2AX foci formation in *RNASEH2A*^{KO}

cells depends on TOP1. Images are representative of five biologically independent experiments. **d**, Wild-type and *RNASEH2A*^{KO} HeLa cells were transfected with non-targeting (siCTRL) or TOP1-targeting (siTOP1) short interfering RNAs (siRNAs). Immunoblot of WCEs, probed for TOP1. Actin was used as a loading control. **e**, Representative micrographs of wild-type and *RNASEH2A*^{KO} HeLa cells transfected with siCTRL or siTOP1 immunostained for γ-H2AX. Scale bars, 10 μm. **f**, Quantification of experiments shown in **e**. Mean number of foci per nucleus per experiment (open circles) with the mean of five biologically independent experiments (red lines). At least 100 cells were analysed per sample in each experiment. **g**, TOP1 depletion alleviates PARPi-induced apoptosis in *RNASEH2A*^{KO} cells. Quantification of cleaved caspase-3⁺ wild-type and *RNASEH2A*^{KO} cells transfected with the indicated siRNAs, with or without talazoparib treatment. Data are mean ± s.d. from three biologically independent experiments normalized to untreated cells. At least 10,000 cells were analysed per sample in each experiment. *P* values in **b**, **f**, **g**, are from unpaired two-tailed *t*-tests. See also Extended Data Figs. 5, 6.

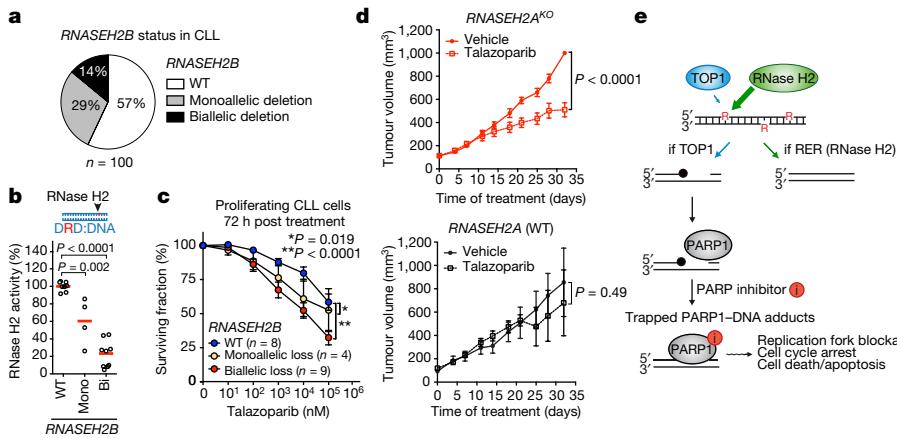


Fig. 4 | Talazoparib selectively suppresses growth of RNase H2 deficient tumours. **a–c**, PARP inhibitors selectively kill RNASEH2B-deficient CLL primary cancer cells. **a**, RNASEH2B deletion frequency in a panel of 100 primary CLL samples, determined by MLPA. **b**, Reduced RNase H2 activity in lysates from CLL samples with monoallelic and biallelic *RNASEH2B* deletions. Top, substrate schematic. Individual data points are the mean of technical duplicates for each sample. Red lines are the mean of individual genotypes ($n=8$ wild type, 4 monoallelic- and 9 biallelic-deleted biologically independent primary CLL samples). Data are normalized to the mean of the wild-type *RNASEH2B* samples. **c**, Reduced survival of CLL cells with monoallelic and biallelic *RNASEH2B* loss following treatment with talazoparib. Individual points are mean ± s.e.m. from $n=8$, 4 and 9 CLL samples as in **b**, each analysed in technical triplicate. *P* values are from an unpaired two-tailed *t*-test (**b**) and two-way

ANOVA (**c**). **d**, Selective inhibition of *RNASEH2A*^{KO} xenograft tumour growth. Wild-type *RNASEH2A* (bottom) or *RNASEH2A*^{KO} (top) HCT116 *TP53*^{KO} cells were injected subcutaneously into bilateral flanks of CD-1 nude mice. Mice were randomized to either vehicle or talazoparib (0.333 mg kg⁻¹) treatment groups ($n=8$ animals per group) and tumour volumes were measured twice-weekly. Data are mean ± s.e.m. *P* values are from a two-way ANOVA. **e**, Model of the processing of genome-embedded ribonucleotides. Genome-embedded ribonucleotides (R) can be processed by TOP1 as an alternative to RNase H2-dependent RER. DNA lesions that engage PARP1 (black circles) are formed as a result, and PARP inhibitors induce PARP1 trapping on these TOP1-dependent lesions, causing replication arrest, persistent DNA damage and cell death. See also Extended Data Figs. 7, 8, Extended Data Table 1, Supplementary Table 3.

to RNase H2 loss as complementation with an *RNASEH2A* transgene abrogated PARPi sensitivity (Extended Data Fig. 8d). We conclude that collateral loss of RNase H2 enhances the vulnerability of cancer cells to PARP-trapping drugs.

Finally, we note that genome-embedded ribonucleotides are by far the most abundant aberrant nucleotides in the genome of cycling cells¹³ and may thus represent a major source of the traps that mediate the cytotoxicity of PARPi alongside base excision repair intermediates. In support of this possibility, *RNASEH2A*^{KO} cells are more sensitive to PARPi than isogenic cell lines with homozygous mutations in the catalytic domain of DNA polymerase-β (*POLB*^{Δ188–190}), a key enzyme in base excision repair (Extended Data Fig. 9). We therefore propose a model whereby the canonical RER pathway and TOP1 compete for the processing of genome-embedded ribonucleotides (Fig. 4e). Whereas RNase H2 cleavage initiates their problem-free removal, the action of TOP1 on ribonucleotides creates PARP-trapping DNA lesions that impair successful completion of DNA replication and the resulting burden of genomic lesions ultimately causes cell death. We propose that the manipulation of genomic ribonucleotide processing could be harnessed for therapeutic purposes and this strategy may expand the use of PARP inhibitors to some tumours that are proficient in homologous recombination.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0291-z>.

Received: 10 July 2017; Accepted: 7 June 2018;

Published online 4 July 2018.

- Lord, C. J. & Ashworth, A. PARP inhibitors: synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017).
- Pommier, Y., O'Connor, M. J. & de Bono, J. Laying a trap to kill cancer cells: PARP inhibitors and their mechanisms of action. *Sci. Transl. Med.* **8**, 362ps17 (2016).
- Hopkins, T. A. et al. Mechanistic dissection of PARP1 trapping and the impact on in vivo tolerability and efficacy of PARP inhibitors. *Mol. Cancer Res.* **13**, 1465–1477 (2015).
- Murai, J. et al. Trapping of PARP1 and PARP2 by clinical PARP inhibitors. *Cancer Res.* **72**, 5588–5599 (2012).
- Cerritelli, S. M. & Crouch, R. J. The balancing act of ribonucleotides in DNA. *Trends Biochem. Sci.* **41**, 434–445 (2016).
- Elstrott, F. et al. BRCA1 mutation analysis of 41 human breast cancer cell lines reveals three new deleterious mutants. *Cancer Res.* **66**, 41–45 (2006).
- Daemen, A. et al. Cross-platform pathway-based analysis identifies markers of response to the PARP inhibitor olaparib. *Breast Cancer Res. Treat.* **135**, 505–517 (2012).
- Wang, G. et al. Identifying drug–gene interactions from CRISPR knockout screens with drugZ. Preprint at <https://www.biorxiv.org/content/early/2017/12/12/232736> (2017).
- Ohle, C. et al. Transient RNA–DNA hybrids are required for efficient double-strand break repair. *Cell* **167**, 1001–1013.e7 (2016).
- Pierce, A. J., Johnson, R. D., Thompson, L. H. & Jasin, M. XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev.* **13**, 2633–2638 (1999).
- Potenski, C. J., Niu, H., Sung, P. & Klein, H. L. Avoidance of ribonucleotide-induced mutations by RNase H2 and Srs2–Exo1 mechanisms. *Nature* **511**, 251–254 (2014).
- Pizzi, S. et al. Reduction of hRNase H2 activity in Aicardi–Goutières syndrome cells leads to replication stress and genome instability. *Hum. Mol. Genet.* **24**, 649–658 (2015).
- Reijns, M. A. et al. Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. *Cell* **149**, 1008–1022 (2012).
- Hiller, B. et al. Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity. *J. Exp. Med.* **209**, 1419–1426 (2012).
- Reijns, M. A. & Jackson, A. P. Ribonuclease H2 in health and disease. *Biochem. Soc. Trans.* **42**, 717–725 (2014).
- Chon, H. et al. RNase H2 roles in genome integrity revealed by unlinking its activities. *Nucleic Acids Res.* **41**, 3130–3143 (2013).
- Kim, N. et al. Mutagenic processing of ribonucleotides in DNA by yeast topoisomerase I. *Science* **332**, 1561–1564 (2011).

- Sparks, J. L. & Burgers, P. M. Error-free and mutagenic processing of topoisomerase 1-provoked damage at genomic ribonucleotides. *EMBO J.* **34**, 1259–1269 (2015).
- Williams, J. S. et al. Topoisomerase 1-mediated removal of ribonucleotides from nascent leading-strand DNA. *Mol. Cell* **49**, 1010–1015 (2013).
- Sekiguchi, J. & Shuman, S. Site-specific ribonuclease activity of eukaryotic DNA topoisomerase I. *Mol. Cell* **1**, 89–97 (1997).
- Pommier, Y., Sun, Y., Huang, S. N. & Nitiss, J. L. Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. *Nat. Rev. Mol. Cell Biol.* **17**, 703–721 (2016).
- Huang, S. N., Williams, J. S., Arana, M. E., Kunkel, T. A. & Pommier, Y. Topoisomerase I-mediated cleavage at unpaired ribonucleotides generates DNA double-strand breaks. *EMBO J.* **36**, 361–373 (2017).
- Koops, T. J. et al. Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Primers* **3**, 16096 (2017).
- Klein, U. et al. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* **17**, 28–40 (2010).
- Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
- Mu, P. et al. SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and Rb1-deficient prostate cancer. *Science* **355**, 84–88 (2017).
- Ku, S. Y. et al. Rb1 and Trp53 cooperate to suppress prostate cancer lineage plasticity, metastasis, and antiandrogen resistance. *Science* **355**, 78–83 (2017).
- Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).

Acknowledgements We thank Y. Pommier and N. Huang for discussions and communication of unpublished results; R. Szilard for critical reading of the manuscript; R. Greenberg for HeLa DR-GFP cells; the IGMM Imaging and Flow Cytometry facilities for assistance and T. Heffernan and N. Feng for providing talazoparib. M.Z. is a Banting postdoctoral fellow. O.M. is supported by an EMBO Long-Term Fellowship (ALTF 7-2015), the European Commission FP7 (Marie Curie Actions, LTFCOFUND2013, GA-2013-609409) and the Swiss National Science Foundation (P2ZHP3_158709). Work in the laboratory of A.P.J. was supported by the Medical Research Council (MRC, U127580972); Work in the laboratory of T.S. was supported by Bloodwise (14031). Work in the laboratories of S.A. and J.M. was supported by grants from the Canadian Cancer Society (#705045; to S.A.) and CIHR (MOP- 142375; to J.M.). Work in the laboratory of J.d.B. was supported by the Movember Foundation, Prostate Cancer UK, the US Department of Defense, the Prostate Cancer Foundation, Stand Up To Cancer, Cancer Research UK, and the UK Department of Health through an Experimental Cancer Medicine Centre grant and work in the laboratory of V.G.B. was supported by Cancer Research UK (grants C157/A25140 and C157/A15703). D.D. is the Thomas Kierans Chair in Mechanisms of Cancer Development and a Canada Research Chair (Tier I) in the Molecular Mechanisms of Genome Integrity. Work in the laboratory of D.D. was funded through CIHR grant FDN143343, Canadian Cancer Society (CCS grants #70389 and #705644), as well as a Grant-in-Aid from the Krembil Foundation.

Reviewer information *Nature* thanks A. Chabes, M. Tarsounas and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.Z. performed the initial CRISPR screens with the help of M.A., A.M., M.Ch., S.A. and J.M.; T.H. analysed the data; M.Z. and O.M. performed suppressor screens and A.M. helped with data analysis. Unless otherwise stated, M.Z. and O.M., with input from M.A.M.R., performed all additional experiments and data analysis. M.A.M.R. performed biochemical characterization of RER-deficient RNase H2, and together with Z.T. and A.F. contributed to the generation of HeLa and HCT116 *RNASEH2A*^{KO} cell lines. A.A., under the supervision of T.S., conducted ex vivo CLL studies and CGH arrays. S.P. and P.M. clinically characterized CLL patients and provided CLL blood samples. R.C. performed MLPA assays. W.Y., M.C. and M.B.L., under the supervision of J.d.B., analysed copy-number alterations (CNAs) in the *RB1–RNASEH2B* region in CRPCs. M.M. and O.M., under the supervision of V.G.B., conducted xenograft experiments. A.P.J. and D.D. designed and directed the study. D.D. and A.P.J. wrote the manuscript with help of M.Z., O.M. and M.A.M.R. and all authors reviewed it.

Competing interests D.D. and T.H. are advisors to Repare Therapeutics.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0291-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0291-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.P.J. or D.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Cell culture. HeLa, RPE1-hTERT and 293T cells were purchased from ATCC and grown in Dulbecco's Modified Eagle Medium (DMEM; Gibco/Thermo Fisher) supplemented with 10% fetal bovine serum (FBS; Wisent), 200 mM GlutaMAX, 1× non-essential amino acids (both Gibco/Thermo Fisher), 100 U ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin (Pen/Strep; Wisent). HCT116 *TP53*^{KO} cells²⁹, a gift from B. Vogelstein, were maintained in modified McCoy's 5A medium (Gibco/Thermo Fisher) supplemented with 10% FBS and Pen/Strep. SUM149PT cells were purchased from Asterand BioScience and grown in a DMEM/F12 medium mixture (Gibco/Thermo Fisher) supplemented with 5% FBS, Pen/Strep, 1 µg ml⁻¹ hydrocortisone and 5 µg ml⁻¹ insulin (both Sigma). Wild-type and *BRCA2*^{KO} DLD-1 cells were purchased from Horizon and maintained in RPMI medium (Gibco/Thermo Fisher) supplemented with 10% FBS and Pen/Strep. All cell lines were grown at 37 °C and 5% CO₂. HeLa, RPE1-hTERT (with the exception of *BRCA1*^{KO} and *POLB*^{Δ188-190} clones) and HCT116 cells were grown at atmospheric O₂. RPE1-hTERT *BRCA1*^{KO} and *POLB*^{Δ188-190} clones, as well as DLD-1 and SUM149PT cell lines were maintained at 3% O₂.

Lentiviral and retroviral transduction. To produce lentivirus, 4.5 × 10⁶ 293T cells in a 10-cm dish were transfected with packaging plasmids (5 µg pVSVg, 3 µg pMDLg/pRRE and 2.5 µg pRSV-Rev) along with 10 µg of transfer plasmid using calcium phosphate. Medium was refreshed 12–16 h later. Virus-containing supernatant was collected ~36–40 h post transfection, cleared through a 0.4-µm filter, supplemented with 4 µg ml⁻¹ polybrene (Sigma) and used for infection of target cells. The TKOv1 library virus was prepared as previously described³⁰. The following antibiotics were used for selection of transductants: puromycin (HeLa, SUM149PT: 2 µg ml⁻¹; RPE1-hTERT 15–20 µg ml⁻¹; each for 48–72 h unless indicated otherwise) and blasticidin (5 µg ml⁻¹, 4–5 days for all cell lines). Cells stably expressing Flag–Cas9–2A–Blast were maintained in the presence of 2 µg ml⁻¹ blasticidin.

To complement the HCT116 *TP53*^{KO} *RNASEH2A*^{KO} cell line, cells were infected with retroviral supernatant produced in amphotropic Phoenix packaging cells³¹ using either pMSCVpuro empty vector (EV) or pMSCVpuro-RNASEH2A-WT in the presence of 4 µg ml⁻¹ polybrene (Sigma) and 48 h later selected for stable integration using 2 µg ml⁻¹ puromycin.

RNASEH2A expression plasmids. A Flag-tagged human *RNASEH2A* cDNA (NM_006397.2; encoding amino acids 2–299) and the D34A/D169A double mutant³² were cloned into the pCW57.1 vector (a gift from D. Root; Addgene #41393) using the Gateway system (Life Technologies/Thermo Fisher) according to the manufacturer's protocol. The P40D and Y210A mutations were generated by site-directed mutagenesis using the following primers (5' to 3'): P40D forward, GGCCCAGCAGTCGCCCTGCCCG; P40D reverse, CGGGCAGG GGCGACGTGCTGGGCC; Y210A forward, GTCTGGGATCATTGGGG GCGCTGAGGCCATAATCAGT; Y210A reverse, ACTGATTATGGCTCAGG CGCCCCAAATGATCCCAAGA. Expression constructs were introduced into HeLa *RNASEH2A*^{KO} cells by lentiviral transduction and expression was induced by the addition of 1 µg ml⁻¹ doxycycline (Clontech) 24 h before starting experiments. The pMSCVpuro-RNASEH2A-WT plasmid was generated by cloning the coding sequence of human *RNASEH2A* into pMSCVpuro-Dest, a Gateway-compatible version of pMSCVpuro (Clontech), and introduced into HCT116 *TP53*^{KO} *RNASEH2A*^{KO} cells by retroviral transduction.

sgRNA target sequences. sgRNAs targeting the following sequences (5' to 3') were used to generate CRISPR knockouts. *RNASEH2A*, TGCCCGCCATCGA CGCCC and CCCGTGCTGGGTGCCCT (for HeLa *RNASEH2A*^{KO}); GACCTATTGGAGAGCGAGC (for HeLa Cas9, RPE1-hTERT, HeLa DR-GFP); *RNASEH2B*, TCCACCACAACTGATCAAG; *PARP1*, TAACGATGTCCA CCAGGCCA; *BRCA1*, AAGGGTAGCTTTAGAAGGC; *POLB*, GAGAACATCC ATGTCACAC; *lacZ*, CCCGAATCTCTATCGTCGG; *PSMD1*-1, TGTGCGCTA CGGAGCTGCAA; *PSMD1*-2, ACCAGAGGCCACAAATAAGCCA; *PSMB2*-1, ATGTTCTTGTGCCCTCCGAC; *PSMB2*-2, AATATTGTCCAGATGAAG GA; *EIF3D*-1, TGTAGTTGCCCTCATGGCC; *EIF3D*-2, AGACGACCTGTCA TCCGCA; *TP53*, CAGAATGCAAGAACGCCAGA.

Vectors expressing the Cas9n D10A nuclease together with guide RNAs designed against exon 1 and intron 1 of human *RNASEH2A* were generated by cloning annealed DNA oligonucleotides into pSpCas9n(BB)-2A-GFP and pSpCas9n(BB)-2A-Puro vectors (Addgene plasmid #48140 and #48141, respectively; gifts from F. Zhang) as previously described³³. All other sgRNA-expressing constructs were generated by cloning annealed DNA oligonucleotides into lentiGuide-Puro or lentiCRISPR v2 vectors (Addgene #52963 and 52961, gifts from F. Zhang) as previously described³⁴.

RNA interference. TOP1 was targeted with 40 nM of either a custom siRNA (siTOP1, target site sequence AAGGACTCCATCAGATACTAT, Sigma) previously described³⁵ or an ON-TARGETplus SMARTpool siRNA (siTOP1-SP, L-005278-00, Dharmacon/BD Technologies), that has previously been used to knock down TOP1^{36–38}. A custom siRNA targeting luciferase (siCTRL,

CTTACGCTGAGTACTTCGA, Sigma) or an ON-TARGETplus non-targeting pool (siCTRL-SP, D-001810-10-05, Dharmacon/BD Technologies) were used as controls³⁹. siRNA oligonucleotides were transfected in Opti-MEM reduced-serum medium using oligofectamine (Life Technologies/Thermo Fisher). To improve knockdown efficiency for the ON-TARGETplus siRNA, a second round of transfection was conducted after 24 h. Following siRNA transfection, cells were seeded either for cell cycle analysis (24 h post last transfection) or for immunofluorescence analysis (48 h post transfection) as described below. Knockdown was optimised to minimize cell death, while maintaining efficient TOP1 depletion (apoptosis levels ≤ 14% of control transfected cells).

DNA damaging drugs. PARP inhibitors olaparib, talazoparib and veliparib were purchased from Selleck Chemicals. Talazoparib for the xenograft experiments was a gift of T. Heffernan and N. Feng. Methyl methanesulfonate (MMS) and aphidicolin were obtained from Sigma. Concentrations and durations of treatment are indicated in the sections below and in the respective figures.

Generation of Cas9-expressing cells. Cells were transduced with the Lenti-FLAG-Cas9-2A-Blast vector³⁰ and transductants were selected with blasticidin. Cells were then seeded at low densities (500–1,000 cells, depending on the cell line) on 15-cm dishes and single colonies were isolated using glass cylinders. Cas9 expression was confirmed by immunoblotting and gene editing efficiency was tested as follows.

Cells were transduced at a low (~0.3) multiplicity of infection (MOI) with either a control *LacZ* sgRNA construct or sgRNA constructs targeting essential genes *PSMD1*, *PSMB2* and *EIF3D*³⁰ and then selected with puromycin. Cells (2.5 × 10⁴) were subsequently seeded in 6-well plates, medium was exchanged 3 days later and the experiment was terminated at day six. Cells were trypsinized, resuspended in medium and the live cell count was determined by trypan blue exclusion on a ViCELL instrument (Beckman Coulter). Cell numbers were plotted relative to *sgLacZ*-transduced samples.

Generation of CRISPR knockout cell lines. To establish HeLa and HCT116 *TP53*^{KO} *RNASEH2A*^{KO} cell lines, 0.5 × 10⁶ cells were seeded in 6-well plates and transfected with two vectors encoding both Cas9n and sgRNAs targeting *RNASEH2A* (derivatives of pSpCas9n(BB)-2A-GFP and pSpCas9n(BB)-2A-Puro) using Lipofectamine 2000 (Life Technologies/Thermo Fisher). Forty-eight hours after transfection, single GFP⁺ cells were sorted into 96-well plates on a BD FACSJazz instrument (BD Biosciences) and grown until colonies formed. *RNASEH2A*^{KO} clones were selected on the basis of the size of PCR amplicons from the targeted region to detect clones that underwent editing, which was subsequently confirmed by Sanger sequencing. The oligonucleotides (5' to 3') used for PCR amplification and sequencing of targeted *RNASEH2A* loci were ACCCGCTCCTGCAGTATTAG and TCCCTTGGTGCAGTCAATC. The absence of functional *RNASEH2A* was confirmed by immunoblotting, an RNase H2 activity assay and alkaline gel electrophoresis as described below. Functionally wild-type *RNASEH2A* clones were identified in parallel and used as controls.

To generate the remaining CRISPR-edited HeLa and RPE1-hTERT cell lines, cells were electroporated with 5 µg of vectors encoding the sgRNA (lentiGuide-Puro, for cells stably expressing Cas9) or encoding both the sgRNA and Cas9 (lentiCRISPR v2) using an Amaxa Nucleofector II instrument (Lonza). RPE1-hTERT cells (0.7 × 10⁶) in a buffer containing 100 mM Na₂HPO₄ (pH 7.75), 10 mM KCl and 11 mM MgCl₂ were electroporated using program T-23. For HeLa cells, the Amaxa Cell Line Nucleofector Kit R (Lonza) was used with program I-13 according to the manufacturer's instructions. Cells were re-plated into antibiotic-free McCoy's 5A medium supplemented with 10% FBS and allowed to recover for 24 h. Puromycin was subsequently added to growth medium to enrich for transfectants and removed 24 h later. Cells were then cultured for an additional 3–5 days to provide time for gene editing and eventually seeded at low densities (400–1,000 cells, depending on cell line) on 15-cm dishes. Single colonies were isolated using glass cylinders two to three weeks later. SUM149PT Cas9 *RNASEH2B*^{KO} cells were generated by transient transfection of parental SUM149PT Cas9 cells with a lentiGuide-puro-sgRNASEH2B construct using Lipofectamine 2000 (Thermo Fisher) as per the manufacturer's protocol (2 µg plasmid DNA and 2 µl of Lipofectamine 2000 was used for 1 × 10⁵ cells in a 6-well plate). Transfected cells were selected with puromycin for 24 h, grown for an additional 4 days and single clones were isolated as above.

Targeted clones were identified by immunofluorescence and/or immunoblotting and successful gene editing was confirmed by PCR and TIDE analysis (<https://tide-calculator.nki.nl>)⁴⁰. The following PCR primers (5' to 3') were used for amplification of targeted loci in *RNASEH2A* forward, AGATCTGGAGGCCCTGAA GT GG, *RNASEH2A* reverse, AGTGGCTGTATCATGTGACAGGG; *RNASEH2B* forward, TAGATGGTGTGCTGTG, *RNASEH2B* reverse, TGCTCAGCTGTCATTGACC; *BRCA1* forward, TCTCAAAGTATT CATTTCCTGG TGCC, *BRCA1* reverse, TGAGCAAGGATCATAAAATGTTGG; *PARP1* forward, AAGCAAACAGGACTGCCAGC, *PARP1* reverse, TACGCCACTGCACTC CAGC; *POLB* forward, TTACTGTTGTCATCA CAGATTCTGC, *POLB*

reverse, AGCAACTCATGGAAGAATAATAGG; *TP53* forward, GCATTGAAGTCTCATGGAAGC; *TP53* reverse, TCACT GCCATGGA GGAGC. **Generation of wild-type and RNASEH2A^{KO} HeLa FUCCI cells.** To establish wild-type RNASEH2A^{KO} HeLa cells expressing the FUCCI cell cycle reporters mKO2-Cdt1 and mAG-Geminin⁴¹, wild-type and RNASEH2A^{KO} HeLa cells were transduced at a low MOI with pLenti6-mKO2-Cdt1 and pLenti6-mAG-Geminin vectors and transductants were selected with 2 µg ml⁻¹ blasticidin. Subsequently, cells positive for both mKO2-Cdt1 and mAG-Geminin fluorescence were collected by sorting on a BD Biosciences FACS Aria II instrument, expanded and used for further experiments. Expression of mKO2-Cdt1 and mAG-Geminin was confirmed by immunofluorescence and FACS analysis.

CRISPR-Cas9 screening. CRISPR screens were performed as described³⁰. Cas9-expressing cells were transduced with the lentiviral TKOv1 library at a low MOI (~0.2–0.3) and puromycin-containing medium was added the next day to select for transductants. Selection was continued until 72 h post transduction, which was considered the initial time point (day 0). At this point the transduced cells were split into technical triplicates. During negative-selection screens (for PARPi sensitizers), cells were subcultured at day 3 and at day 6 each of the three replicates was divided into two populations. One was left untreated and to the other a dose of olaparib amounting to 20% of the lethal dose (LD₂₀) (HeLa, 2 µM; RPE1-hTERT, 0.5 µM; SUM149PT, 0.2 µM) was added. Cells were grown with or without olaparib until day 15 and subcultured every three days. Sample cell pellets were frozen at each time point for genomic DNA (gDNA) isolation. A library coverage of ≥200 cells per sgRNA was maintained at every step. Positive-selection screens (for suppressors of sensitivity) were carried out in a similar way, but the untreated control was left out, an LD₈₀ dose of talazoparib was used (20 and 50 nM for HeLa and RPE1-hTERT, respectively), cells were subcultured only once after drug addition (day 12–13) and screens were terminated at day 18. Library coverage was ≥100 cells per sgRNA.

gDNA from cell pellets was isolated using the QIAamp Blood Maxi Kit (Qiagen) and genome-integrated sgRNA sequences were amplified by PCR using the KAPA HiFi HotStart ReadyMix (Kapa Biosystems). i5 and i7 multiplexing barcodes were added in a second round of PCR and final gel-purified products were sequenced on Illumina HiSeq2500 or NextSeq500 systems to determine sgRNA representation in each sample. DrugZ⁸ was used to identify gene knockouts, which were depleted from olaparib-treated day-15 populations but not depleted from untreated cells. Gene knockouts enriched at day 18 as compared to day 6 in positive-selection screens were identified using MAGeCK⁴².

Gene Ontology and interaction network analyses. PANTHER (<http://pantherdb.org>)⁴³ was used to identify Gene Ontology (GO) biological processes enriched in datasets of screen hits as compared to genome-wide representation. Hits (FDR ≤ 0.01 in at least one cell line and FDR ≤ 0.1 in at least two cell lines) from individual cell lines or hits common to at least two cell lines were analysed with the 'statistical overrepresentation test' (Gene Ontology ontology database released 28 February 2017; annotation dataset 'GO biological process complete') with Bonferroni correction for multiple testing. Mapping of the hits on the HumanMine protein interaction network was done using the esyN interface (<http://www.esyn.org/>). The network was then exported and visualized in Cytoscape v.3.4.0 (<http://www.cytoscape.org/>) and the node sizes adjusted to be proportional to the averaged DrugZ score over the three cell lines.

Clonogenic survival assays. To determine PARPi sensitivity cells were seeded on 10-cm dishes (500–3,000 cells per plate, depending on cell line and genotype) into drug-free medium or media containing a range of PARPi concentrations. Cells were either treated for 2 days with talazoparib followed by additional 9–12 days of growth in drug-free media (HeLa, SUM149PT), treated for 7 days with talazoparib followed by 5–6 days in drug-free media (RPE1-hTERT, HCT116), or treated continuously for 12–13 days with olaparib. The cultures were incubated at 3% O₂ with the exception of the experiment in Fig. 2j, which was carried out at atmospheric O₂. Medium (with or without PARPi) was refreshed every 4–7 days in all cases. At the end of the experiment medium was removed, cells were rinsed with PBS and stained with 0.4% (w/v) crystal violet in 20% (v/v) methanol for 30 min. The stain was aspirated and plates were rinsed 2 × in ddH₂O and air-dried. Colonies were manually counted and data were plotted as surviving fractions relative to untreated cells. To calculate EC₅₀ values the data were fit to a three-parameter dose response model (log(inhibitor) versus normalized response) using the nonlinear regression function in Graphpad PRISM v6.0.

To analyse the synthetic lethality of combined *BRCA1* and RNASEH2B knock-outs, wild-type *BRCA1* and *BRCA1*^{KO} RPE1-hTERT Cas9 *TP53*^{KO} cells were transduced at a high MOI (>1.0) with lentiGuide sgRNA constructs targeting either RNASEH2B or *LacZ* (control) and seeded for clonal growth 48 h later. Wild-type and *BRCA1*^{KO} colonies were grown at 3% O₂ for 12 and 20 days (owing to the slower growth of *BRCA1*-deficient cells), respectively. Synthetic lethality between RNase H2 and *BRCA2* was assessed by transducing wild-type and *BRCA2*^{KO} DLD-1 cells with either an empty lentiCRISPR v2 vector or lentiCRISPR v2

constructs carrying sgRNASEH2A or sgRNASEH2B. Cells were selected with puromycin and seeded for clonogenic assays 7 days post infection. Clones were grown at 3% O₂ for 11 (wild type) or 14 days (*BRCA2*^{KO}).

Immunofluorescence microscopy. To analyse γ-H2AX focus formation, cells were grown on coverslips for 24 h, incubated in medium containing 10 µM EdU for 20 min to label cells undergoing DNA replication, then pre-extracted for 5 min on ice with ice-cold buffer (25 mM HEPES, pH 7.4, 50 mM NaCl, 1 mM EDTA, 3 mM MgCl₂, 300 mM sucrose and 0.5% Triton X-100) and fixed with 4% paraformaldehyde (PFA) for 15 min at room temperature (RT). After fixation, cells were washed with PBS and blocked in 3% FBS in PBS for 30 min at room temperature. EdU immunolabelling was performed using the Click-iT EdU Imaging Kit (Invitrogen, C10337). Afterwards cells were incubated with a primary mouse antibody against γ-H2AX (Millipore 05-636; 1:800) for 1.5 h at room temperature and then stained with anti-mouse secondary antibodies conjugated to Alexa Fluor 568 (Life Technologies) for 1 h at room temperature. Coverslips were mounted using Vectashield antifade mounting medium with 4,6-diamidino-2-phenylindole (DAPI; Vector Laboratories). For quantification of γ-H2AX foci images were visualized on a Zeiss AxioPlan 2 microscope with a 40× Plan-neofluar objective, captured using Micro-Manager (<https://open-imaging.com/>) and analysed using an ImageJ-based script as previously described⁴⁴. Nuclei were defined on the basis of DAPI staining, and γ-H2AX foci were detected using the 'Find maxima' function of ImageJ within each nuclear region. Exposure time, binning, microscope settings, light source intensity and the noise level in the 'Find maxima' function were kept constant for all the samples within each individual experiment. More than 100 cells were analysed per condition in each experiment.

For combined γ-H2AX and RAD51 immunofluorescence, 0.25 × 10⁶ cells were seeded on coverslips and ~24 h later were subjected to 3 Gy of X-ray irradiation. Two-, four- or six-hour post irradiation cells were incubated with nuclear extraction buffer (20 mM HEPES pH 7.4, 20 mM NaCl, 5 mM MgCl₂, 0.5% NP-40, 1 mM DTT and protease inhibitors) for 10 min on ice, rinsed with ice-cold PBS and subsequently fixed with 4% PFA for 10 min at room temperature. Cells were blocked in immunofluorescence blocking buffer (10% goat serum, 0.5% NP-40, 0.5% saponin in PBS) for 30 min and incubated with primary antibodies diluted in blocking buffer (Santa-Cruz Biotechnologies rabbit anti-RAD51 and Millipore mouse anti-γ-H2AX; 1:150 and 1:2,000, respectively) for 2 h at room temperature. Cells were then washed with PBS (3 × 5 min) and stained with fluorescent secondary antibodies (Alexa Fluor 488-conjugated goat anti-rabbit IgG and Alexa Fluor 555-conjugated goat anti-mouse IgG, Life Technologies/Thermo Fisher; 1:1,000 in blocking buffer) and 0.5 µg ml⁻¹ DAPI for 1 h at room temperature. Cells were washed as above, mounted in ProLong Gold mounting medium (Life Technologies/Thermo Fisher) and imaged using a Zeiss LSM780 laser-scanning microscope. Cells with more than five colocalizing γ-H2AX and RAD51 foci were quantified by manual counting. At least 100 cells per condition were analysed in each experiment.

Immunofluorescence and flow cytometry. For detection of apoptotic cells by cleaved caspase-3 immunofluorescence FACS, 0.5 × 10⁶ cells were plated on 6-cm dishes and either left untreated or treated with PARPi for 48 to 72 h (PARPi doses are indicated in respective figures). For analysis of apoptotic cells following TOP1 depletion, 0.25 × 10⁶ cells were plated on 6-cm dishes, transfected with siCTRL or siTOP1 the next day and 24 h post-transfection were either left untreated or treated with PARPi for 48 h. Medium was removed and stored in a conical tube, cells were collected by trypsinization, resuspended in the original conditioned medium and centrifuged at 524g for 5 min at 4°C. Pellets were washed in PBS and fixed in 1 ml 4% PFA for 10 min at room temperature. Cells were pelleted as above, resuspended in 100 µl PBS and chilled on ice. 900 µl of –20°C methanol was then added drop-wise while gently vortexing. Fixed cells were stored at –20°C overnight or longer.

Before staining, cells were spun down as above, washed in PBS and blocked in immunofluorescence blocking buffer (see the 'Immunofluorescence' section). Cells were then centrifuged and resuspended in 200 µl of diluted rabbit anti-cleaved caspase-3 antibody (Cell Signaling #9661; 1:800 in immunofluorescence blocking buffer). After 2 h incubation the antibody was diluted with 2 ml PBS, cells were spun down, and incubated for 1 h in 200 µl Alexa Fluor 488-conjugated goat anti-rabbit secondary antibody (Molecular Probes/Thermo Fisher, 1:1,000 in IF blocking buffer). The antibody was diluted with 2 ml PBS, cells were centrifuged, resuspended in 1 ml PBS and cleaved caspase-3 signal was analysed on BD FACSCalibur or BD LSRFortessa X-20 instruments. Data were analysed using FlowJo software (Tree Star). See Supplementary Fig. 2 for examples of gating strategies.

For analysis of recovery from talazoparib-induced replication blockage, cells were treated with 1 µM talazoparib for 24 h, washed extensively with PBS and grown in drug-free medium for additional 10, 24 or 34 h. Cells were then collected, fixed, stained as described above using an anti-γ-H2AX primary antibody (JBW301, Millipore #05-636, 1:1,000 in blocking buffer) and finally DNA was labelled with propidium iodide (see below).

Cell cycle analysis by FACS. Cells (0.5×10^6) were seeded on 6-cm dishes into medium with or without PARPi (doses and durations are indicated in respective figures). Cells were then collected by trypsinization, resuspended in medium and centrifuged (233g, 5 min, 4°C). Pellets were resuspended in PBS, centrifuged again and resuspended in 1 ml propidium iodide staining buffer (20 $\mu\text{g ml}^{-1}$ propidium iodide, 0.02% Triton X-100, 0.2 mg ml^{-1} RNase A in PBS). Cells were stained for 15 min at 37°C and analysed on a BD FACSCalibur or BD LSR Fortessa X-20 instruments.

For combined propidium iodide and EdU staining, cells were treated and collected as above and fixed in 70% ethanol (added dropwise while gently vortexing) overnight at -20°C. Cells were then centrifuged as above, washed in PBS and incubated with 10 μM Alexa Fluor 488 azide (Molecular Probes/Thermo Fisher) in a buffer containing 100 mM Tris-HCl pH 8.5, 1 mM CuSO₄ and 100 mM ascorbic acid for 30 min before centrifugation, washing in PBS and propidium iodide staining. See Supplementary Fig. 2 for examples of gating strategies.

Sister chromatid exchange assay. HeLa cells (0.5×10^6) were seeded in 10-cm dishes and bromodeoxyuridine (BrdU) (final concentration 10 μM) was added the next day. BrdU containing medium was refreshed 24 h later and cells were grown for another 22 h (46 h BrdU incubation in total). 100 ng ml^{-1} KaryoMAX colcemid (Gibco/Thermo Fisher) was added for the final 2 h and cells were collected as follows.

Growth medium was removed and stored in a conical tube. Cells were gently washed with 1 ml of trypsin (the trypsin wash was combined with the original medium), trypsinized, resuspended in the original conditioned medium (with a trypsin wash) and centrifuged (233g, 5 min, 4°C). Cells were then washed with PBS, spun down, resuspended in pre-warmed 75 mM KCl and incubated for 30 min at 37°C. Cells were centrifuged again, the supernatant was removed and cells were fixed by drop-wise addition of 1 ml fixative (ice-cold methanol: acetic acid, 3:1) while gently vortexing. An additional 10 ml of fixative was then added and cells were fixed at 4°C for at least 16 h. Once fixed, metaphases were dropped on glass slides, rinsed with fixative and air-dried overnight (protected from light).

To visualize sister chromatid exchanges (SCE) slides were rehydrated in PBS for 5 min and stained with 2 $\mu\text{g ml}^{-1}$ Hoechst 33342 (Molecular Probes/Thermo Fisher) in 2 × SSC (final 300 mM NaCl, 30 mM sodium citrate, pH 7.0) for 15 min. Stained slides were placed in a plastic tray, covered with a thin layer of 2 × SSC and irradiated with 254 nM UV light (~5400 J m⁻²). Slides were subsequently dehydrated in a 70%, 95% and 100% ethanol series (5 min each), air-dried and mounted in DAPI-containing ProLong Gold mounting medium (Molecular Probes/Thermo Fisher). Images were captured on a Zeiss LSM780 laser-scanning microscope.

DR-GFP assay and quantitative image-based cytometry. HeLa DR-GFP cells (a gift from R. Greenberg) were transduced with either a lentiCRISPR v2 empty vector or sgRNA-expressing constructs targeting *RNASEH2A* or *RNASEH2B*. Seven days after transductions 4–5 $\times 10^3$ cells were plated per well of 96-well imaging plates (Corning 3603) and next day either mock transfected or transfected with either 100 ng of a plasmid expressing I-SceI or a GFP-expressing plasmid (to assess transfection efficiency) using Lipofectamine 2000. Medium was exchanged 6–8 h later and at 48 h post-transfection cells were fixed in 4% paraformaldehyde. Immunofluorescence for RNASEH2A was performed as described above and plates were imaged on an InCell Analyzer 6000 automated microscope (GE Life Sciences) with a 20 × objective. Image analysis was performed using Columbus (PerkinElmer). Nuclei were segmented and a sum of DAPI intensity, mean RNASEH2A intensity and mean GFP intensity was quantified for each nucleus. Cells showing a DNA content between 1N and 2N were selected based on DAPI intensity, RNASEH2A⁺ and RNASEH2A⁻ populations were separated and percentages of GFP⁺ cells were calculated. Only RNASEH2A⁺ cells were analysed in vector-infected samples, whereas only RNASEH2A⁻ cells were considered in sgRNA-transduced samples. Percentages of GFP⁺ cells in each sample were normalized to the transfection efficiency of a control GFP plasmid.

Immunoblotting. Cell pellets were resuspended in hot 2 × sample buffer (166.7 mM Tris-HCl pH 6.8, 2% SDS, 20 mM DTT, 10% glycerol, 0.01% bromophenol blue) at a concentration of 5 $\times 10^6$ cells ml^{-1} and denatured at 95°C for 5 min. An equivalent of 0.25–1 $\times 10^5$ cells was separated by SDS-PAGE and transferred to a nitrocellulose or PVDF (for RNASEH2B) membrane. Membranes were blocked with 5% milk in TBST (TBS + 0.1% Tween-20) for at least 1 h at room temperature and incubated with primary antibodies diluted in 5% milk in TBST overnight at 4°C. Membranes were then washed three times with TBST, incubated with horseradish peroxidase-conjugated secondary antibodies for 1 h at room temperature, washed again and protein bands were detected using the SuperSignal West Pico enhanced chemiluminescence reagent (Thermo Fisher).

To assess the efficiency of TOP1 depletion, WCEs were obtained by lysis and sonication of cells in UTB buffer (8 M urea, 50 mM Tris-HCl, pH 7.5, 150 mM β -mercaptoethanol, protease inhibitor cocktail (Roche)). WCEs for RNase H activity assays and for determining protein levels of the RNase H subunits were prepared by lysing cells in 50 mM Tris-HCl pH 8.0, 280 mM NaCl, 0.5% NP-40,

0.2 mM EDTA, 0.2 mM EGTA, 10% glycerol (v/v), 1 mM DTT and 1 mM PMSF for 10 min on ice, and subsequent addition of an equal volume of 20 mM HEPES pH 7.9, 10 mM KCl, 1 mM EDTA, 10% glycerol (vol/vol), 1 mM DTT and 1 mM PMSF for an additional 10 min. WCEs were cleared by centrifugation (17,000g for 10 min at 4°C) and protein concentration was determined using the Bradford assay (Protein Assay Kit, BioRad). Protein samples (35 μg total protein) were run on NuPAGE 4–12% Bis-Tris Protein Gels (Thermo Fisher Scientific) and transferred to nitrocellulose or PVDF membranes. Membranes were blocked in 5% milk in TBST and immunoblotting was performed as described above.

Immunoprecipitation. Cells were collected by trypsinization, washed once with PBS supplemented with 1 μM ADP-HPD (PARG inhibitor; Enzo) and 4 $\times 10^6$ cells were snap-frozen in liquid nitrogen and then lysed in 1 ml of lysis buffer (50 mM HEPES pH 8.0, 100 mM KCl, 2 mM EDTA, 0.5% NP-40, 10% glycerol, 1 mM DTT, complete protease inhibitor cocktail (Roche), 1 μM ADP-HPD). Lysates were incubated with gentle rotation at 4°C for 15 min and then centrifuged at 15,000g for 10 min. Fifty microlitres of total cell lysates were used as input and 950 μl were incubated with 5 μl of mouse anti-PARP1 antibody (Enzo (F1-23) ALX-804-211-R050) for 5 h at 4°C. Protein G-agarose beads (60- μl slurry; Pierce) were added for an additional hour. Beads were collected by centrifugation, washed four times with lysis buffer and eluted by boiling in 60 μl 2 × sample buffer. Samples were run on an 8% SDS-PAGE gel and immunoblotting was performed as described above (see 'Immunoblotting' section).

To analyse PARP1 poly(ADP-ribosylation) in a specific phase of the cell cycle, wild-type and *RNASEH2A*^{KO} HeLa FUCCI cells were trypsinized, washed once with PBS, collected in tubes with PBS supplemented with 3% FCS and 1 μM ADP-HPD, and sorted on the basis of mKO2-Cdt1 (G1 phase) and mAG-Geminin (S/G2/M phases) fluorescence on a BD Biosciences FACS Aria II instrument. 4 $\times 10^6$ FACS-sorted cells were snap-frozen and lysed as described above. Equivalent amounts of proteins (~0.5–1 mg) were incubated with 25 μl of PARP1-Trap_A pre-equilibrated bead slurry (ChromoTek) for 2.5 h at 4°C, washed four times with lysis buffer and eluted by boiling in 2 × sample loading buffer (31.25 mM Tris pH 6.8, 25% glycerol, 1% SDS, 0.01% bromophenol blue, β -mercaptoethanol) before immunoblotting. Samples were run on a NuPAGE 4–12% Bis-Tris Protein Gel (Thermo Fisher Scientific) and immunoblotting was performed as described above.

Antibodies. The following antibodies were used for immunofluorescence and immunoblotting at the indicated dilutions: sheep anti-pan-RNase H2 (raised against human recombinant RNase H2¹³, immunoblots 1:1,000, immunoprecipitation 5 μl per 1 ml lysate); rabbit anti-RNASEH2C (Proteintech 16518-1-AP; immunoblots 1:1,000); rabbit anti-RNASEH2A (Origene TA306706, immunoblots 1:1,000); mouse anti-RNASEH2A (Abcam ab92876; immunofluorescence 1:500); mouse anti-RNASEH2A G-10 (Santa Cruz Biotechnologies sc-515475; western blot 1:1,000); mouse anti- γ H2AX JBW301 (Millipore 05-636, immunofluorescence 1:800–1:2,000); rabbit anti-RAD51 H-92 (Santa Cruz Biotechnologies sc-8349, immunofluorescence 1:150); rabbit anti-BRCA1⁴⁵ (immunoblots 1:1,000); mouse anti-Cas9 (Diagenode C15200203, immunoblots 1:1,000); rabbit anti-PARP1 H-250 (Santa Cruz Biotechnologies sc-7150, immunoblots 1:1,000); mouse anti-PAR 10H (Enzo ALX-804-220-R100, immunoblots 1:1,000); rabbit anti-Topoisomerase I (Abcam ab109374; immunoblots 1:5,000); rabbit anti-DYKD-DDDK (Cell Signaling Technologies 2368, immunoblots 1:1,000); rabbit anti-actin (Sigma A2066, immunoblots 1:5,000); mouse anti- α -tubulin DM1A (Millipore CP06, immunoblots 1:5,000); mouse anti- α -tubulin B512 (Sigma T6074, immunoblots 1:5,000); rabbit anti-GAPDH (Sigma G9545, immunoblots 1:20,000); mouse anti-vinculin (Sigma V9264, immunoblots 1:1,000); rabbit anti-DNA polymerase beta (Abcam ab26343, immunoblots 1:1,000); rabbit anti-cleaved caspase-3 (Cell Signaling Technologies 9661S, immunofluorescence 1:800).

Cell Titer Glo assay. Two hundred cells per condition were plated on 96-well assay plates in technical triplicates either in drug-free medium or in a range of MMS concentrations. MMS was washed out 24 h later and cells were grown in drug-free medium for another 48 h. Cell viability was analysed using the Cell Titer Glo assay kit (Promega) according to the manufacturer's instructions. Luminescence was read on an Envision 2104 plate reader (Perkin Elmer).

Detection of ribonucleotides in genomic DNA. Total nucleic acids were isolated from 10⁶ cells by lysis in ice-cold buffer (20 mM Tris-HCl pH 7.5, 75 mM NaCl, 50 mM EDTA) and subsequent incubation with 200 $\mu\text{g ml}^{-1}$ proteinase K (Roche) for 10 min on ice followed by addition of sarkosyl (Sigma) to a final concentration of 1%. Nucleic acids were sequentially extracted with TE-equilibrated phenol, phenol:chloroform:isoamyl alcohol (25:24:1), and chloroform, and then precipitated with isopropanol. Nucleic acids were collected by centrifugation, washed with 75% ethanol, air-dried and dissolved in nucleic acid-free water.

For alkaline gel electrophoresis, 500 ng of total nucleic acids were incubated with 1 pmol of purified recombinant human RNase H2³² and 0.25 μg of DNase-free RNase (Roche) for 30 min at 37°C in 100 μl reaction buffer (60 mM KCl, 50 mM Tris-HCl pH 8.0, 10 mM MgCl₂, 0.01% BSA, 0.01% Triton X-100). Nucleic acids

were ethanol-precipitated, dissolved in nuclease-free water and separated on a 0.7% agarose gel in 50 mM NaOH, 1 mM EDTA. After electrophoresis, the gel was neutralized in 0.7 M Tris-HCl pH 8.0, 1.5 M NaCl and stained with SYBR Gold (Invitrogen). Imaging was performed on a FLA-5100 imaging system (Fujifilm), and densitometry plots generated using an AIDA Image Analyzer (Raytest).

RNase H2 activity assay. Recombinant RNase H2 was expressed in Rosetta-2 *Escherichia coli* cells using a polycistronic construct based on pGEX6P1 (pMAR22) and purified as previously described³². Site-directed mutagenesis to introduce the D34A and D169A or P40D and Y210A mutations was performed using the Quikchange method (Agilent). To measure enzyme activity, a range of RNase H2 concentrations (0.06–2 nM) was incubated with 2 μ M substrate in 5- μ l reactions (in a buffer containing 60 mM KCl, 50 mM Tris-HCl pH 8.0, 10 mM MgCl₂, 0.01% BSA and 0.01% Triton X-100) at 37 °C for 30 min or 1 h. Substrate was formed by annealing a 3'-fluorescein-labelled oligonucleotide (GATCTGAGCCTGGGaGCT, DRD-DNA; or gaucugagccugggagcu, RNA-DNA; uppercase DNA, lowercase RNA) to a complementary 5'-dabcyl-labelled DNA oligonucleotide (Eurogentec). Reactions were stopped by adding an equal volume of 96% formamide, 20 mM EDTA, and heating at 95 °C. Products were resolved by denaturing PAGE (20%, 1 \times TBE), visualized on a FLA-5100 imaging system (Fujifilm) and quantified using ImageQuant TL (GE Healthcare).

To assess RNase H2 activity in WCEs a FRET-based fluorescent substrate release assay was performed as previously described³². RNase H2 specific activity was determined against a DRD-DNA substrate (described above). Activity against a double-stranded DNA substrate of the same sequence was measured and used to correct for non-RNase H2 activity against the DRD-DNA substrate. Reactions were performed in 100 μ l of buffer with 250 nM substrate in 96-well flat-bottomed plates at 25 °C. Whole cell lysates were prepared as described above, and the final protein concentration used per reaction was 100 ng μ l⁻¹. Fluorescence was read for 100 ms using a VICTOR2 1420 multilabel counter (Perkin Elmer), with a 480-nm excitation filter and a 535-nm emission filter.

Ex vivo CLL studies. Peripheral blood mononuclear cells were isolated from blood samples collected from patients with a new or existing diagnosis of CLL, irrespective of the stage of disease or duration or type of treatment from two Birmingham hospitals (Heartlands and Queen Elizabeth). This study was approved by the South Birmingham Ethics Committee (REC number 10/H1206/58), performed according to institutional guidelines and written consent was obtained from all participants.

Primary CLL cells (5×10^4) and CD40L-expressing mouse embryonic fibroblasts (5×10^3) were seeded in each well of a 96-well plate (Corning) in 100 μ l of RPMI medium supplemented with 10% FBS (Sigma-Aldrich, UK) and 25 ng ml⁻¹ IL-21 (eBioscience)⁴⁶. After 24 h, 200 μ l more medium was gently added and cells were incubated for another 48 h. The medium was then aspirated, replaced with 200 μ l of medium containing talazoparib and cells were incubated for a further 72 h. The cytotoxic effect of PARPi was determined by propidium iodide exclusion as measured by flow cytometry with an Accuri C6 flow cytometer (Applied Biosystems). Only cells which entered the cell cycle upon stimulation (as determined by forward- and side-scatter FACS profiles), were considered for analysis. Data was expressed as a surviving fraction relative to untreated cells. For gating strategies, see Supplementary Fig. 2.

MLPA assay. Genomic DNA was isolated from primary CLL cells using the Flexigene kit (Qiagen). To identify deletions in *RNASEH2B* gene the MLPA assay was performed on approximately 100 ng of genomic DNA (gDNA) per sample using the P388-A2 SALSA MLPA kit (MRC-Holland) according to the manufacturer's protocol. Two microlitres of amplified products were separated by capillary electrophoresis on an ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems) with a GeneScan 600 LIZ Size Standard (Thermo Fisher). Data were analysed using GeneMarker software v.2.4.0 (SoftGenetics). Data were normalized using gDNA from four control reference samples. Copy number changes represented as a MLPA ratio were detected by comparing normalized peak intensities between the reference and the CLL samples. The MLPA ratio thresholds (X) were set as follows: $0.75 \geq X \leq 1.25$, diploid sample; $0.4 \geq X < 0.75$, monoallelic deletion; $X < 0.4$, biallelic deletion. Samples showing either a s.d. of control probes above 0.15, or samples with large Q fragment peaks and with more than four control probes having MLPA ratios out of diploid range were excluded from the analysis.

CGH array. Genotyping of CLL samples was accomplished using HumanCoreExome BeadChip arrays (Illumina) by UCL Genomics in accordance with the Infinium HTS Assay protocol (15045736_A, Illumina). Genotypes were called by GenomeStudio software Genotyping Module v.3.1 (Illumina). A call rate of 98% was accepted as the primary quality control for each sample. Log R Ratio and B Allele Frequency values generated by the GenomeStudio software were used to assess allelic losses in chromosome 13q.

Analysis of CNAs in the RB1–RNASEH2B region in CRPC. CRPC ($n = 226$) whole exome sequencing data generated by the International Stand Up To Cancer/Prostate Cancer Foundation Prostate Cancer Dream Team were downloaded and re-analysed^{28,47}. Paired-end sequencing reads were aligned to the human reference

genome (GRCh37/hg19) using BWA (v.0.5.9), with default settings and re-aligned using stampy (v.1.0.2). ASCAT (v.2.3) was used to estimate CNA, tumour purity and ploidy.

Xenograft experiments. Female athymic CD-1 nude mice (5–7 weeks old, Charles River Laboratories) were used for in vivo xenograft studies and quarantined for at least 1 week before experiments. Exponentially growing wild-type *RNASEH2A* or *RNASEH2A*^{KO} HCT116 *TP53*^{KO} cells were injected subcutaneously into the bilateral flanks of each animal (2×10^6 cells per flank). Tumours were measured by caliper every 3–4 days and tumour volume was determined by the formula (length \times width²) $/2$. When the tumour volumes reached approximately 100 mm³ (10 days after injection), mice were randomized into treatment and control groups (8 animals per group, 32 animals in total; sample size was determined based on previous relevant studies). Talazoparib (BMN673, 0.333 mg kg⁻¹, pharmacological grade, a gift from T. Heffernan and N. Feng) or vehicle (10% *N,N*-dimethylacetamide (ACROS Organics), 5% Solutol HS 15 (Sigma-Aldrich) in PBS (Gibco)) was administered once daily by oral gavage (0.1 ml per 10 g of body weight) for the indicated length of time, or until the tumour reached the maximum size (15 mm in any direction) or ulcerated, or a body conditioning score of two was reached, as determined by UK Home Office regulations. The data reported are the average tumour volumes per mouse. Individual flanks that showed no evidence of tumour growth before initiation of treatment were excluded from subsequent measurements and analysis.

A subsequent experiment was performed by injecting exponentially growing HCT116 *TP53*^{KO} *RNASEH2A*^{KO} cells complemented either with an empty vector or a vector encoding wild-type *RNASEH2A* (2×10^6 cells per flank). To increase the potential treatment window, mice were randomized into treatment and control groups (8 animals per group, 32 animals in total), and treatment started 3 days after injection when palpable tumours were formed. The treatment was administered as described above. Animals that showed no evidence of tumour growth on both flanks within the first 11 days of treatment were excluded from analysis.

The technician performing tumour measurements was blinded to the experimental design and identity of cells injected. All animal studies were carried out under Project Licence PPL 70/8897 approved by the UK Home Office and by the University of Edinburgh Animal Welfare and Ethical Review Body.

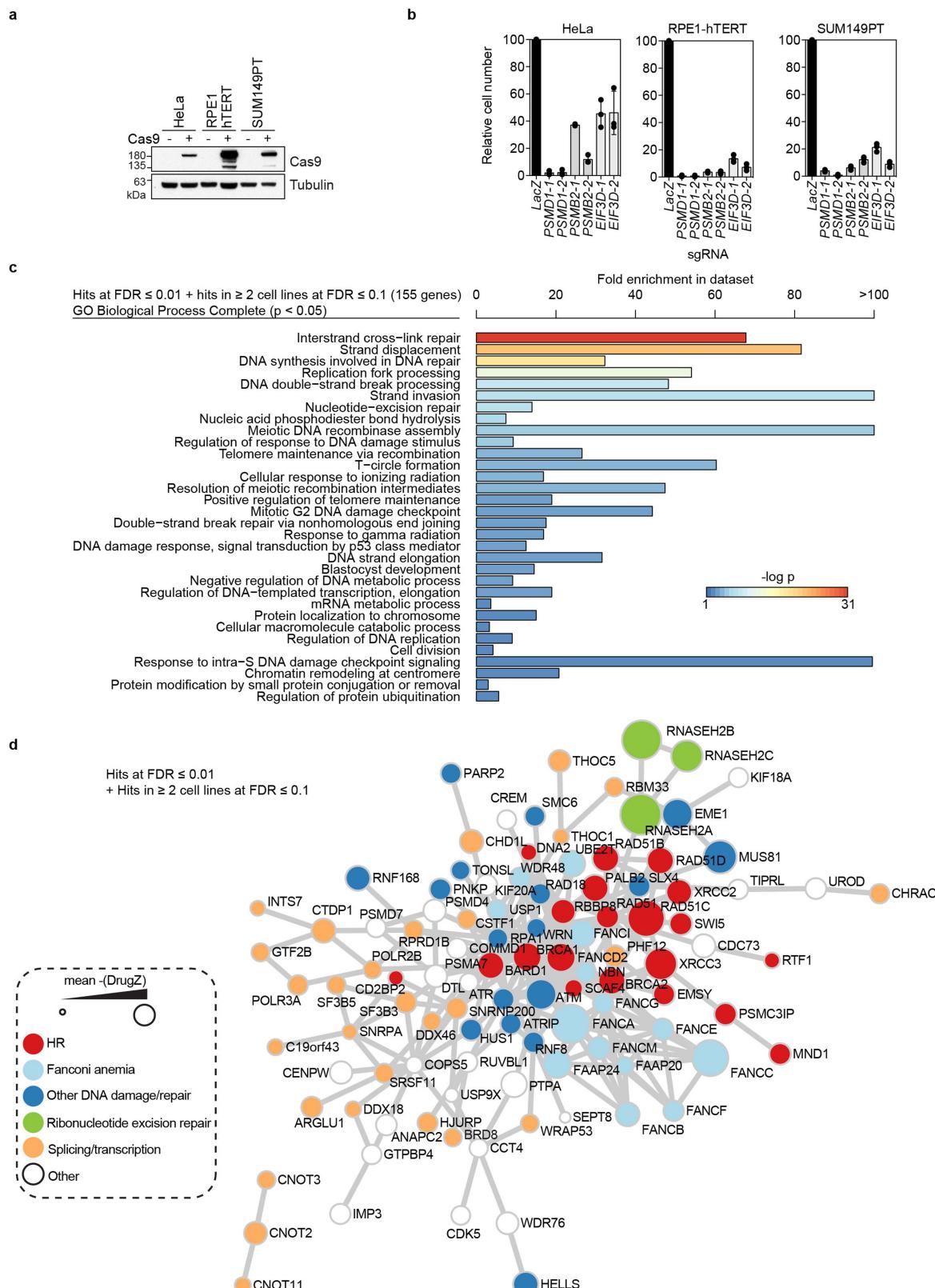
Statistical analysis. Data were analysed using a two-tailed Student's *t*-test and a two-way ANOVA under the assumption of normal distribution for biological parameters. No corrections for multiple testing were made. Tests used are indicated in respective figure legends. The number of samples (n) in figure legends represents independent biological replicates, unless stated otherwise. No statistical methods were used to determine the sample size before starting experiments. Cell biology experiments were not randomized and the investigators were not blinded with regards to sample allocation and evaluation of the experimental outcome. For xenograft experiments blinding and randomization were performed.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Results of PARP inhibitor CRISPR screens, source data for mouse xenograft experiments, unprocessed images of immunoblots and examples of gating strategies for FACS experiments are provided as Supplementary Information. All other datasets generated during this study are available from the corresponding authors upon reasonable request.

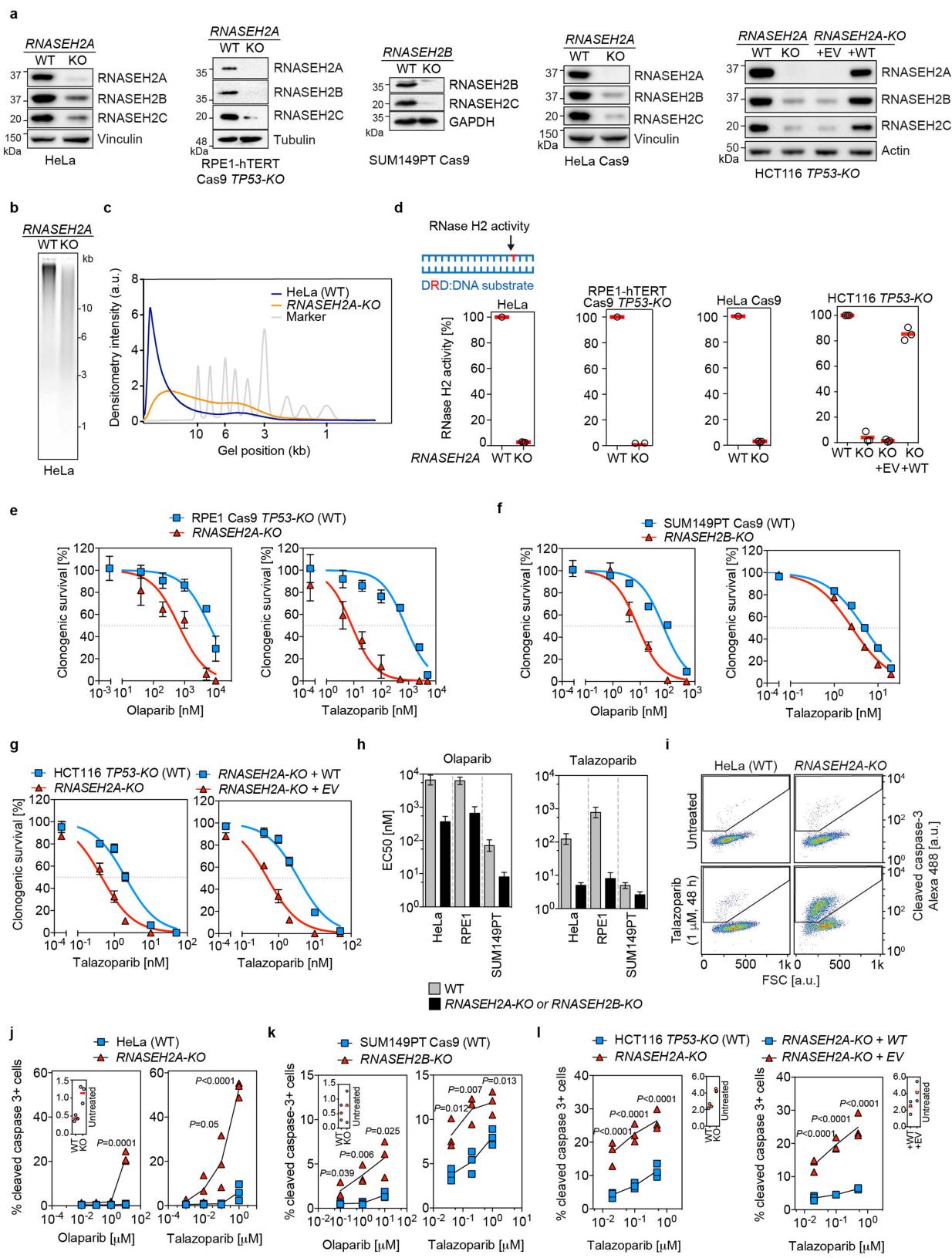
29. Bunz, F. et al. Requirement for p53 and p21 to sustain G2 arrest after DNA damage. *Science* **282**, 1497–1501 (1998).
30. Hart, T. et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).
31. Swift, S., Loresen, J., Achacoso, P. & Nolan, G. P. Rapid production of retroviruses for efficient gene delivery to mammalian cells using 293T cell-based systems. *Curr Protoc Immunol* **31**, 10.17.14–10.17.29 (2001).
32. Reijns, M. A. et al. The structure of the human RNase H2 complex defines key interaction interfaces relevant to enzyme function and human disease. *J. Biol. Chem.* **286**, 10530–10539 (2011).
33. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protocols* **8**, 2281–2308 (2013).
34. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
35. Solier, S. et al. Genome-wide analysis of novel splice variants induced by topoisomerase I poisoning shows preferential occurrence in genes encoding splicing factors. *Cancer Res.* **70**, 8055–8065 (2010).
36. Naughton, C. et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat. Struct. Mol. Biol.* **20**, 387–395 (2013).
37. Helmrich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* **44**, 966–977 (2011).
38. Puc, J. et al. Ligand-dependent enhancer activation regulated by topoisomerase I activity. *Cell* **160**, 367–380 (2015).
39. Harley, M. E. et al. TRAIP promotes DNA damage response during genome replication and is mutated in primordial dwarfism. *Nat. Genet.* **48**, 36–43 (2016).

40. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014).
41. Sakaue-Sawano, A. et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008).
42. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
43. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377–D386 (2013).
44. Reynolds, J. J. et al. Mutations in DONSON disrupt replication fork stability and cause microcephalic dwarfism. *Nat. Genet.* **49**, 537–549 (2017).
45. Escribano-Díaz, C. et al. A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Mol. Cell* **49**, 872–883 (2013).
46. Pascutti, M. F. et al. IL-21 and CD40L signals from autologous T cells can induce antigen-independent proliferation of CLL cells. *Blood* **122**, 3010–3019 (2013).
47. Robinson, D. et al. Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
48. Murai, J. et al. Stereospecific PARP trapping by BMN 673 and comparison with olaparib and rucaparib. *Mol. Cancer Ther.* **13**, 433–443 (2014).



Extended Data Fig. 1 | CRISPR screens for determinants of PARPi sensitivity. This figure is related to Fig. 1. **a**, Cas9 immunoblot of WCEs from parental HeLa, RPE1-hTERT and SUM149PT cells and clones stably transduced with a lentiviral Flag–Cas9–2A–Blast construct. Tubulin was used as a loading control. The immunoblot is representative of two biologically independent experiments. **b**, Validation of CRISPR–Cas9 gene editing efficiency in Cas9-expressing HeLa, RPE1-hTERT and SUM149PT clones. Cell proliferation was monitored after transduction with a control sgRNA construct (sgLacZ) or sgRNAs targeting essential genes *PSMD1*,

PSMB2 and *EIF3D*³⁰. Solid circles, individual values. Data are mean \pm s.d. from three technical replicates normalized to sgLacZ. **c**, Gene Ontology terms significantly ($P < 0.05$, binomial test with Bonferroni correction) enriched among hits from olaparib screens common to at least two cell lines. Enrichment was analysed using PANTHER. **d**, esyN network analysis of interactions between hits common to at least two cell lines. Node size corresponds to mean DrugZ score across cell lines. 77 out of 155 genes are mapped on the network.

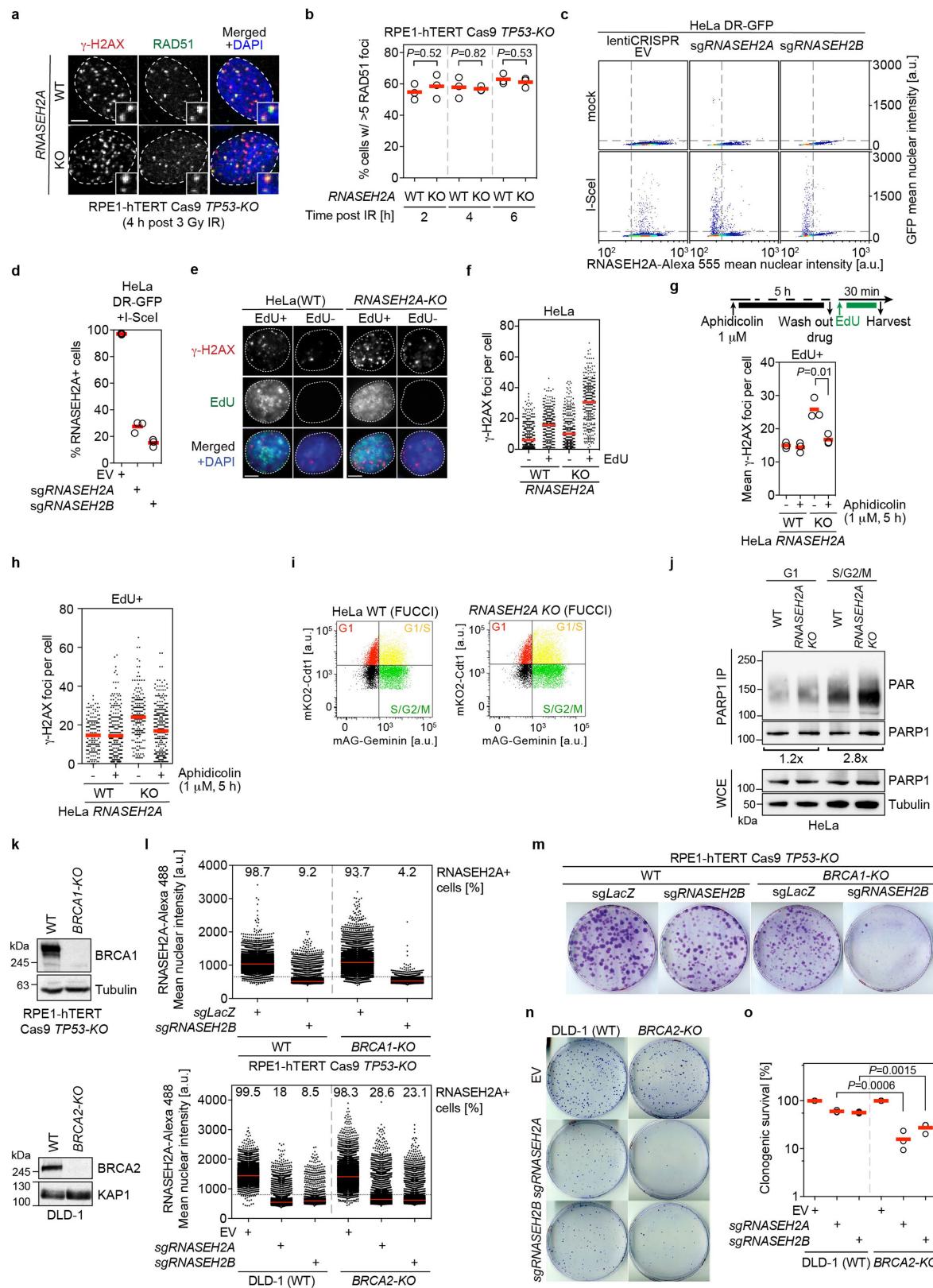


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | RNase H2 deficiency leads to PARPi sensitivity.

This figure is related to Fig. 2a, b. **a**, CRISPR-mediated inactivation of *RNASEH2A* or *RNASEH2B* in the cell lines used in this manuscript. WCEs of indicated cell lines and genotypes were processed for immunoblotting using antibodies against RNASEH2A, RNASEH2B or RNASEH2C. Vinculin, tubulin and GAPDH were used as loading controls. Representative immunoblots of at least two biologically independent experiments. **b-d**, Abolished RNase H2 enzymatic activity and increased levels of genome-embedded ribonucleotides in *RNASEH2A*^{KO} cells. **b**, Analysis of total nucleic acids from wild-type and *RNASEH2A*^{KO} HeLa cells treated with recombinant RNase H2 and separated by alkaline agarose gel electrophoresis. Ribonucleotide-containing genomic DNA from *RNASEH2A*^{KO} HeLa cells is nicked and therefore has increased electrophoretic mobility¹³. Data are representative of three biologically independent experiments. **c**, Densitometric quantification of the alkaline gel shown in **b**. **d**, Cleavage of an RNase H2-specific double-stranded DNA oligonucleotide with a single incorporated ribonucleotide (DRD:DNA; ribonucleotide position is shown in red) by wild-type and *RNASEH2A*^{KO} WCEs of the indicated cell types was measured using a fluorescence quenching-based assay³². Data are individual values (open circles) with the mean (red lines) of three biologically independent experiments.

e-l, RNase H2 deficiency leads to PARPi sensitivity in multiple cell types. **e-g**, Clonogenic survival assays of the indicated cell lines treated with the indicated PARPi. Data are mean \pm s.d. from three biologically independent experiments normalized to untreated cells. Solid lines show a nonlinear least-squares fit of the data to a three-parameter dose-response model. **h**, EC₅₀ values for olaparib (left) and talazoparib (right) in the indicated cell lines as determined by nonlinear least-squares fitting of the data in **e-g** and Fig. 2a, b. Data are EC₅₀ values \pm 95% confidence intervals. **i-l**, Increased apoptosis in *RNASEH2A*^{KO} HeLa, Cas9 *RNASEH2B*^{KO} SUM149PT and *RNASEH2A*^{KO} HCT116 cells following PARPi treatment. **i**, Cleaved caspase-3 immunofluorescence and flow cytometry profiles of untreated and talazoparib-treated wild-type and *RNASEH2A*^{KO} HeLa. FSC, forward scatter. **j-l**, Percentages of cleaved caspase-3⁺ (caspase-3⁺) cells of the indicated genotypes treated with the indicated PARPi. Data are individual values (coloured symbols) with the mean (solid lines) of three biologically independent experiments. Insets and outsets, levels of cleaved caspase-3⁺ cells without PARPi treatment. Red lines, mean of three biologically independent experiments. *P* values are from unpaired two-tailed *t*-tests. In **a**, **d**, **g**, **l**, *RNASEH2A*^{KO} HCT116 cells were transduced either with an empty vector (+EV) or a full-length RNASEH2A expression construct (+WT).

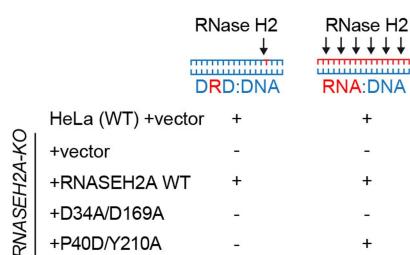


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Functional homologous recombination, increased replication-associated DNA damage and synthetic lethality with the loss of BRCA1 or BRCA2 in RNase H2-deficient cells. This figure is related to Fig. 2. **a–d**, Homologous recombination is not affected by the inactivation of RNase H2. **a**, Micrographs of wild-type and *RNASEH2A*^{KO} RPE1-hTERT Cas9 *TP53*^{KO} and cells exposed to 3 Gy of X-rays (IR) and assessed using γ -H2AX and RAD51 immunofluorescence 4 h later. Images are representative of three biologically independent experiments. **b**, Quantification of the experiment in **a** at the indicated time points after irradiation, plotted as a percentage of cells with more than five γ -H2AX and RAD51 colocalizing foci. Data are individual values (open circles) and the mean (red lines) from three biologically independent experiments. *P* values, unpaired two-tailed *t*-test. **c**, Quantitative image-based cytometry (QIBC) plots of DR-GFP experiments in Fig. 2e. Each point shows the mean GFP and RNASEH2A immunofluorescence intensities per nucleus of mock- or I-SceI-transfected HeLa DR-GFP cells transduced with indicated Cas9 sgRNA constructs (EV = empty vector). Dashed lines separate RNASEH2A⁺ and RNASEH2A⁻ and GFP⁺ and GFP⁻ cell populations. Data are representative of three biologically independent experiments. **d**, Quantification of RNASEH2A⁺ cells in DR-GFP experiments shown in **c** and Fig. 2e as determined by QIBC. Data are individual values (open circles) and the mean (red lines) of three biologically independent experiments. **e–h**, Replication-dependent endogenous DNA damage in RNase H2-deficient cells. **e**, Micrographs for experiments quantified in Fig. 2g. γ -H2AX immunofluorescence in EdU⁺ and EdU⁻ wild-type and *RNASEH2A*^{KO} HeLa cells. Scale bars, 5 μ m. **f**, Quantification of γ -H2AX foci per nucleus in experiments shown in **e** and Fig. 2g. Dots, foci number in individual nuclei; red lines, mean from three biologically independent experiments. **g, h**, Wild-type and *RNASEH2A*^{KO} HeLa cells were treated with aphidicolin and EdU as indicated in the schematic (**g**, top), and immunostained with γ -H2AX antibodies. Mean number of foci per EdU⁺ nucleus in each experiment (**g**, open circles) or the number of foci in individual EdU⁺ nuclei (**h**, dots) were quantified. Red lines are the mean from three biologically independent experiments.

independent experiments; at least 100 cells were analysed per sample in each experiment. *P* value, unpaired two-tailed *t*-test. **i, j**, Increased poly(ADP-ribosylation) of PARP1 in G1 as well as in S/G2/M phases in *RNASEH2A*^{KO} cells. **i**, FACS plots of wild-type and *RNASEH2A*^{KO} HeLa cells expressing the FUCCI cell cycle reporters mKO2-Cdt1 and mAG-Geminin⁴¹. Data are representative of two biologically independent experiments. **j**, PARP1 immunoprecipitates from WCEs of FUCCI-sorted G1 or S/G2/M wild-type and *RNASEH2A*^{KO} HeLa cells, probed with the indicated antibodies. Images are representative of two biologically independent experiments. Tubulin was used as a loading control. Densitometric quantification of PAR signals normalized to immunoprecipitated PARP1 is shown as fold changes from wild-type to *RNASEH2A*^{KO} cells. **k–o**, Inactivation of RNase H2 in *BRCA1*- or *BRCA2*-deficient backgrounds results in synthetic lethality. **k**, *BRCA1* and *BRCA2* expression, respectively, in wild-type and *BRCA1*^{KO} RPE1-hTERT *TP53*^{KO} cells (top) or wild-type and *BRCA2*^{KO} DLD-1 cells (bottom). WCEs were processed for immunoblotting with the indicated antibodies. Tubulin and KAP1 were used as loading controls. Immunoblots are representative of at least two biologically independent experiments. **l**, RNase H2 levels in cells used in **m–o** and Fig. 2i. Cells were transduced with the indicated sgRNA vectors and processed for RNASEH2A immunofluorescence. Each point represents mean RNASEH2A intensity per nucleus as measured by QIBC (*n* = 1 experiment). At least 2,000 cells were analysed per sample. Percentages of RNASEH2A⁺ cells in individual samples are shown above each plot. **m**, Clonogenic survival assays quantified in Fig. 2i. Images are representative of three biologically independent experiments. **n, o**, Synthetic lethality after inactivation of *RNASEH2A* or *RNASEH2B* in *BRCA2*-deficient cells. Clonogenic survival of wild-type and *BRCA2*^{KO} DLD-1 cells was assessed after transduction with indicated Cas9 sgRNA vectors. **n**, Representative images of three biologically independent experiments. **o**, Quantification of the experiments shown in **n**. Data are individual values (open circles) with the mean (red lines) of three biologically independent experiments. *P* values are from an unpaired two-tailed *t*-test.

a



HeLa (WT) +vector + +

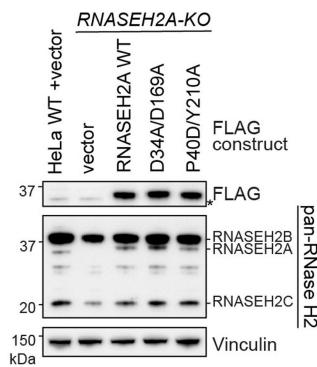
+vector - -

+RNASEH2A WT + +

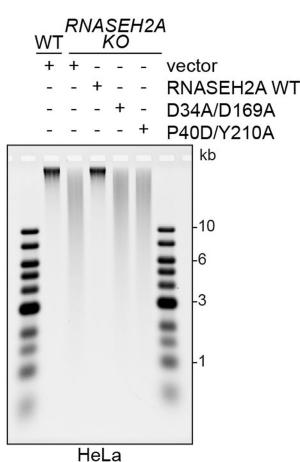
+D34A/D169A - -

+P40D/Y210A - +

b

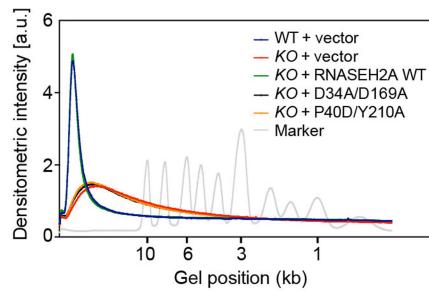


c

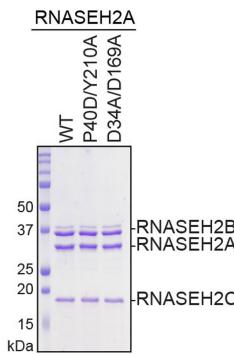


HeLa

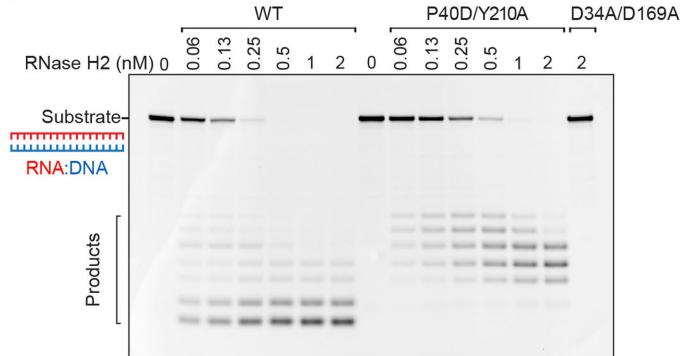
d



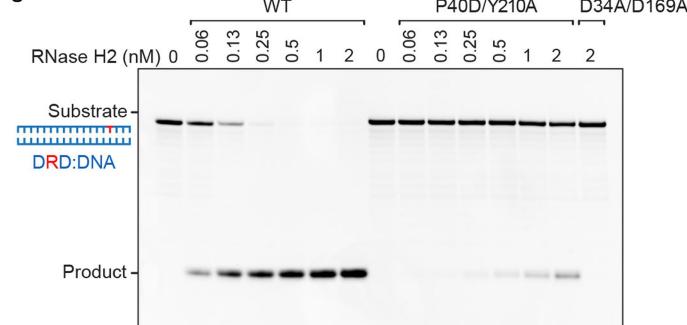
e



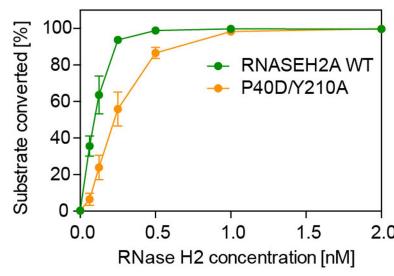
f



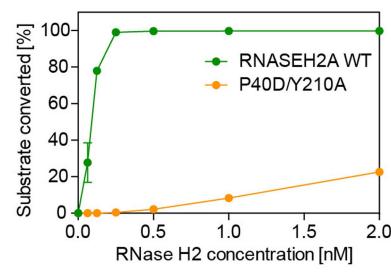
g



h



k

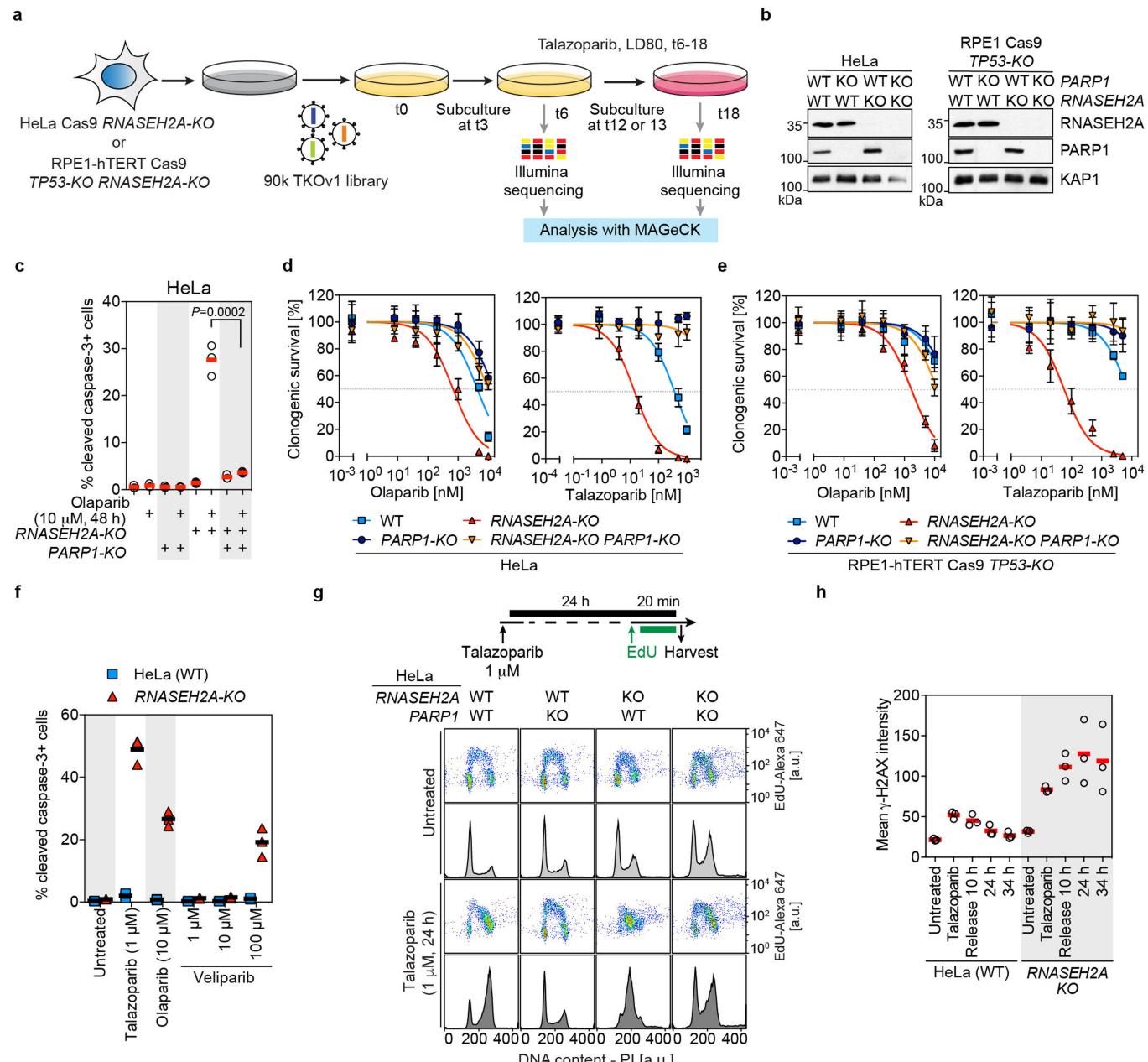


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | A separation-of-function mutant of RNase H2.

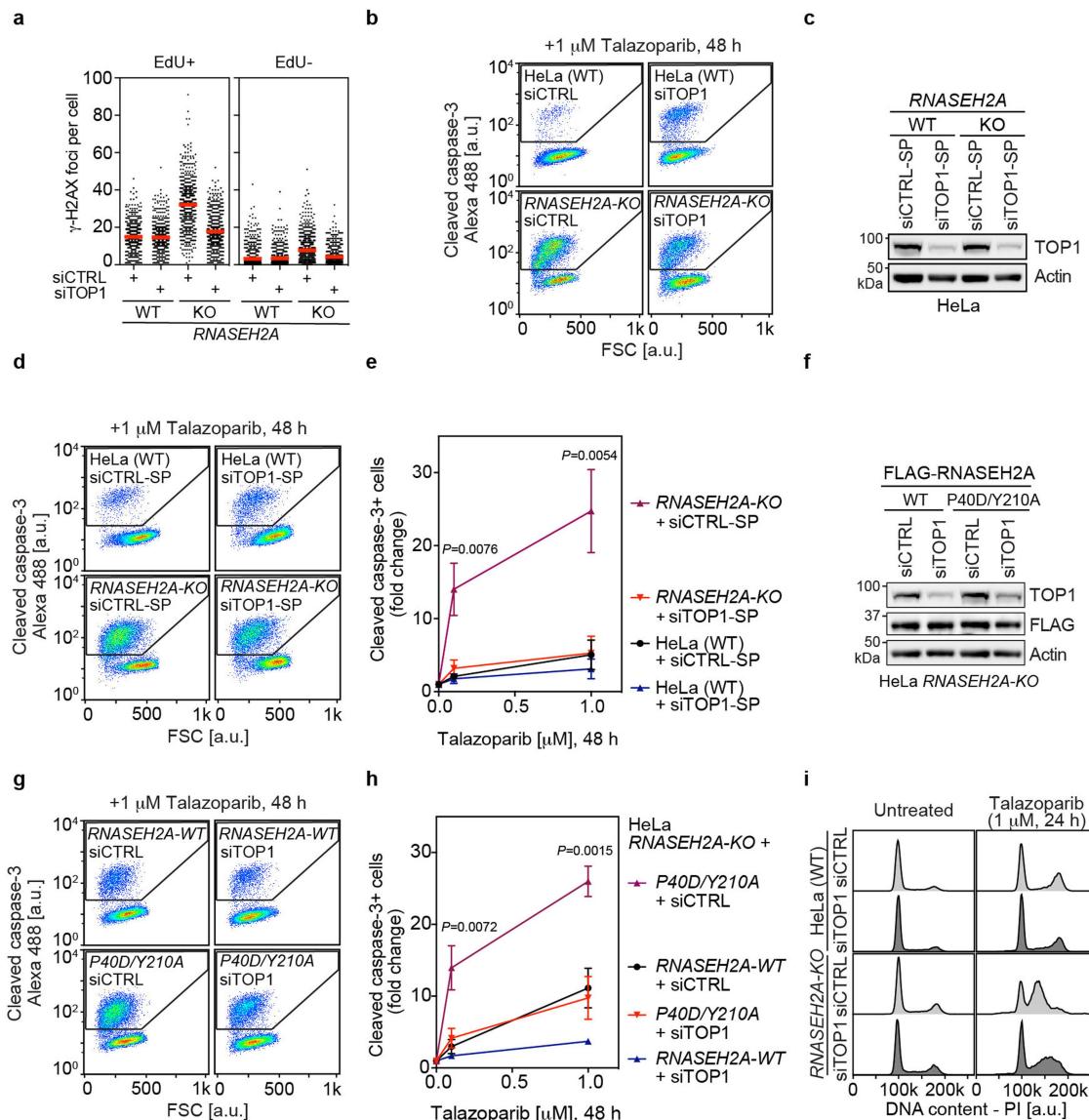
This figure is related to Fig. 2j. RNASEH2A(P40D/Y210A) is a separation-of-function mutant that cannot excise single DNA-embedded ribonucleotides, but cleaves RNA–DNA heteroduplexes (similar to the yeast *rnh201-P45D-Y219A* mutant¹⁶). **a**, Schematic depicting enzymatic activity against two different RNase H2 substrates (DRD–DNA, double-stranded DNA with an embedded ribonucleotide, or RNA–DNA hybrids) in cell lines used in **b–d** and Fig. 2j. Wild-type and *RNASEH2A*^{KO} cells were transduced with either an empty vector or the indicated RNASEH2A constructs. **b**, Complementation of HeLa *RNASEH2A*^{KO} cells with Flag-tagged RNASEH2A variants restores RNase H2 complex protein levels. WCEs from wild-type and *RNASEH2A*^{KO} HeLa and *RNASEH2A*^{KO} cells stably expressing the indicated lentiviral constructs were processed for immunoblotting with the indicated antibodies. Vinculin was used as a loading control. Asterisk indicates a non-specific band. Immunoblots are representative of three biologically independent experiments. **c, d**, Complementation of HeLa *RNASEH2A*^{KO} cells with wild-type RNASEH2A, but not with the D34A/D169A (catalytically dead) or P40D/

Y210A (separation-of-function) mutants, rescues increased levels of genome-embedded ribonucleotides. **c**, Total nucleic acids from the cell lines shown in **a, b** were treated with recombinant RNase H2 and separated by alkaline agarose gel electrophoresis. Image is representative of four experiments. **d**, Densitometric quantification of alkaline gel shown in **c**. **e**, Purified human RNase H2 complexes consisting of RNASEH2B, RNASEH2C and RNASEH2A wild type, P40D/Y210A or D34A/D169A subunits separated by SDS–PAGE and stained with Coomassie blue. One experiment was performed. **f–k**, RNase H2 activity assays with fluorescein-labelled RNA–DNA substrate (**f**) or double-stranded DNA with a single incorporated ribonucleotide (DRD–DNA) (**g**) and increasing amounts of recombinant wild-type, P40D/Y210A or D34A/D169A RNase H2. Products were separated by polyacrylamide gel electrophoresis and detected by fluorescence imaging. Images are representative of three biologically independent experiments. **h, k**, Quantification of gels from **f, g**. Product signal is plotted relative to substrate signal per lane. Data are mean \pm s.d. from three biologically independent experiments.



Extended Data Fig. 5 | PARP1 trapping is the underlying cause of PARPi sensitivity in RNase H2-deficient cells This figure is related to Fig. 3a-c. **a**, Schematic representation of CRISPR screens for suppressors of talazoparib sensitivity in RNase H2-deficient cells. Cas9-expressing cells were transduced with the TKOv1 library, talazoparib was added on day 6 (t6; HeLa, 20 nM; RPE1-hTERT, 50 nM) and cells were cultured in its presence until day 18 (t18). Cells were subcultured once at day 12 (RPE1) or 13 (HeLa). sgRNA representations in the initial (t6) and final (t18) populations were quantified by next-generation sequencing. Gene knockouts that were enriched at day 18 over day 6 were identified by MAGECK⁴². **b**, CRISPR-mediated inactivation of RNASEH2A and/or PARP1 in cell lines used in **c-e** and Fig. 3b. WCEs were processed for immunoblotting with the indicated antibodies. KAP1 was used as a loading control. Immunoblots are representative of two biologically independent experiments. **c-e**, Loss of PARP1 restores PARPi-resistance in RNASEH2A^{KO} cells. **c**, Percentage of cleaved caspase-3⁺ HeLa cells of indicated genotypes with or without olaparib treatment measured by flow cytometry (FACS). Data are individual values (open circles) with the mean (red lines) of three biologically independent experiments. *P* value is from an unpaired two-tailed *t*-test. **d**, **e**, Clonogenic survival assays with HeLa (**d**) and RPE1-hTERT (**e**) cells of the indicated genotypes treated with

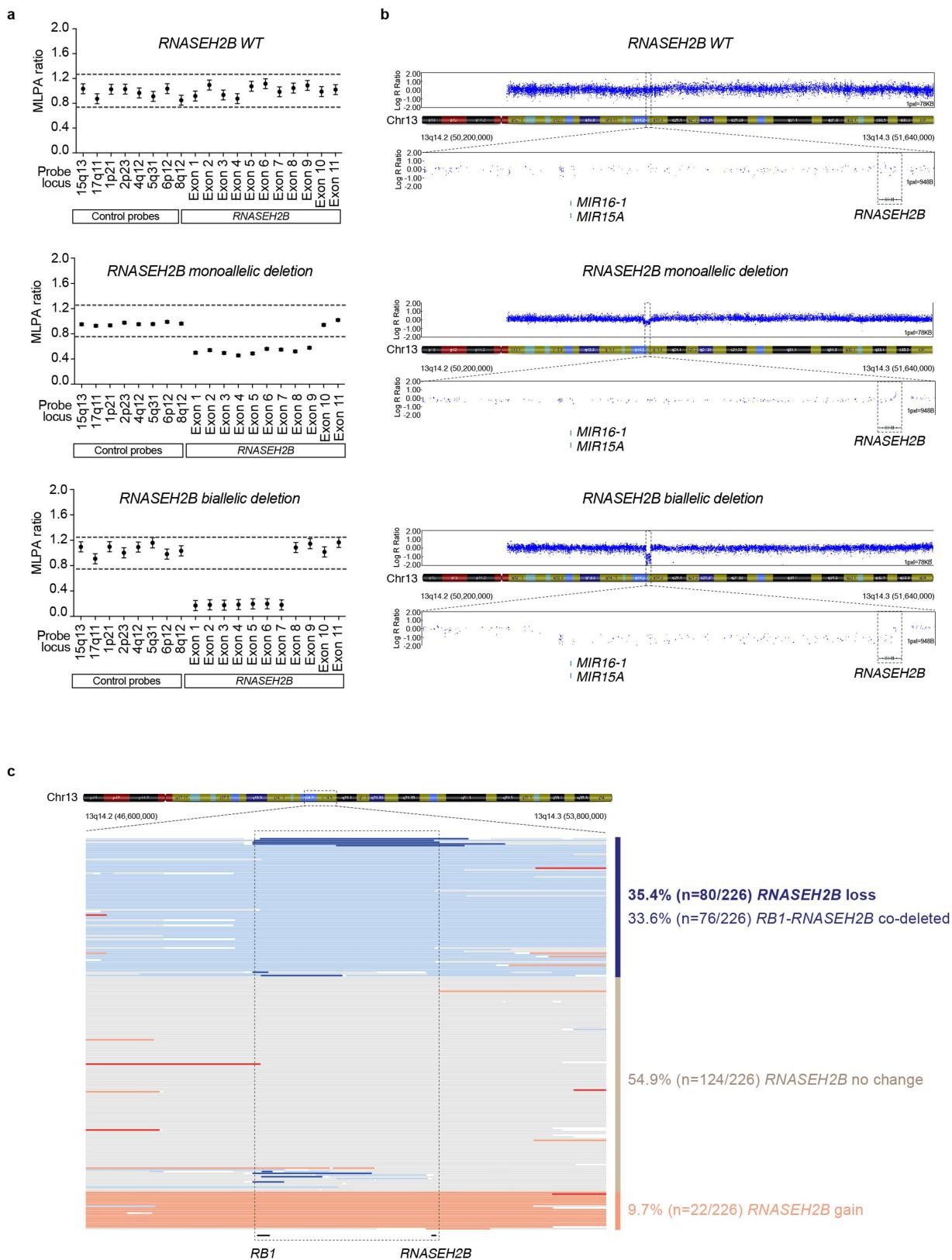
olaparib (left) or talazoparib (right). Data are mean \pm s.d. from three biologically independent experiments. Solid lines show a nonlinear least-squares fit to a three-parameter dose-response model. **f**, Trapping activity of PARPi correlates with the ability to induce apoptosis in RNASEH2A^{KO} cells. Quantification of cleaved caspase-3⁺ wild-type and RNASEH2A^{KO} HeLa cells without treatment or treated with the indicated PARPi. Data are individual values with the mean (black lines) of three biologically independent experiments. Note that PARP-trapping activity decreases as follows: talazoparib > olaparib > veliparib^{4,48}. **g**, PARPi-induced S-phase arrest in RNASEH2A^{KO} cells is alleviated in the absence of PARP1. Top, schematic of talazoparib and EdU treatment. Bottom, EdU (pseudocolor plots) and DNA content (histograms) FACS profiles of untreated and talazoparib-treated wild-type, PARP1^{KO}, RNASEH2A^{KO} and PARP1^{KO}RNASEH2A^{KO} HeLa cells. DNA content was determined by propidium iodide (PI) staining. Data are representative of three biologically independent experiments. **h**, Quantification of mean γ -H2AX intensities in experiments shown in Fig. 3c. Data are individual values (open circles) with the mean (red lines) of three biologically independent experiments. At least 10,000 cells were analysed per sample in each experiment.



Extended Data Fig. 6 | TOP1-mediated cleavage at genome-embedded ribonucleotides leads to PARPi sensitivity in RER-deficient cells.

This figure is related to Fig. 3d–g. **a**, Reduced endogenous DNA damage in TOP1-depleted *RNASEH2A*^{KO} cells. Quantification of γ -H2AX foci per nucleus in the experiments shown in Fig. 3e, f. Dots, foci number in individual nuclei; red lines, mean of five biologically independent experiments. **b–i**, TOP1 depletion alleviates PARPi-induced apoptosis and S-phase arrest in HeLa *RNASEH2A*^{KO} cells (**b–e**) and in *RNASEH2A*(P40D/Y210A) separation-of-function mutant cells (**f–h**). **b**, Cleaved caspase-3 FACS plots for experiments quantified in Fig. 3g. Data are representative of three biologically independent experiments. **c**, Wild-type and *RNASEH2A*^{KO} HeLa cells were transfected with non-targeting (siCTRL-SP) or TOP1-targeting (siTOP1-SP) SMARTpool siRNAs. WCEs analysed by immunoblotting with antibodies to TOP1 and actin (loading control). Images are representative of three biologically independent experiments. **d**, FACS plots of cleaved caspase-3 in wild-type and *RNASEH2A*^{KO} HeLa cells transfected with siCTRL-SP or siTOP1-SP

after talazoparib treatment. **e**, Quantification of the experiment shown in **d**. **f**, *RNASEH2A*^{KO} HeLa cells stably expressing the indicated Flag-tagged constructs were transfected with non-targeting (siCTRL) or TOP1-targeting (siTOP1) siRNAs. WCEs were analysed by immunoblotting with TOP1, Flag and actin (loading control) antibodies. Immunoblots are representative of three biologically independent experiments. **g**, FACS plots of cleaved caspase-3 in *RNASEH2A*^{KO} HeLa cells transfected with siCTRL or siTOP1 and expressing wild-type *RNASEH2A* or the P40D/Y210A mutant. Data are representative of three biologically independent experiments. **h**, Quantification of the experiment shown in **g**. Data in **e**, **h**, are mean \pm s.d. from three biologically independent experiments normalized to untreated cells. At least 10,000 cells were analysed per sample in each experiment. *P* values are from an unpaired two-tailed *t*-test. **i**, Cell cycle profiles, before and after talazoparib treatment, of wild-type and *RNASEH2A*^{KO} HeLa cells transfected with the indicated siRNAs. DNA content was assessed by propidium iodide staining and FACS. Data are representative of three biologically independent experiments.

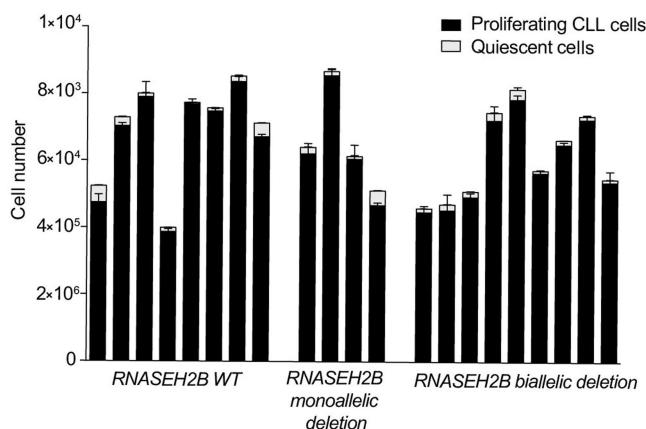


Extended Data Fig. 7 | See next page for caption.

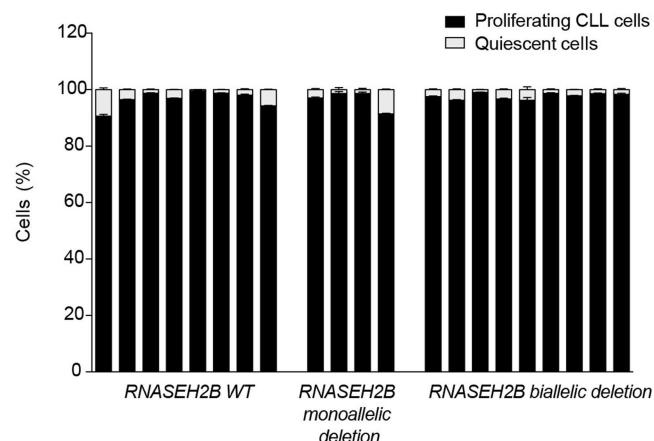
Extended Data Fig. 7 | Frequent collateral loss of *RNASEH2B* in CLL and metastatic CRPC. This figure is related to Fig. 4a–c. **a, b**, MLPA analysis (**a**) and CGH array profiles for chromosome 13q (**b**) of representative CLL samples carrying two wild-type *RNASEH2B* alleles (top), a monoallelic *RNASEH2B* deletion (middle) or biallelic deletion (bottom). **a**, For MLPA analysis, genomic DNA from reference and experimental samples was analysed using probes targeting control loci and individual *RNASEH2B* exons (exon 1–11). The MLPA ratio was calculated per probe and normalized to control probes and reference samples. Error bars indicate s.d. of the mean from eight control probes for each sample. Dashed lines indicate the threshold set for diploid copy number. **b**, For each CGH array profile the *y*-axes of the top and bottom plots indicate copy number probe intensity ($\log R$ ratio) and the *x*-axes mark the

position on chromosome 13 represented by the ideogram (middle). An enlargement of the frequently deleted 13q14.2–14.3 region, including the *miRNA-15A/16-1* gene cluster and the *RNASEH2B* gene, is shown in the bottom plot. One experiment was performed. **c**, *RNASEH2B* is frequently co-deleted with *RB1* in CRPC. Copy number alterations (CNA) in the *RB1–RNASEH2B* region in CRPC ($n = 226$ cases) are shown. Horizontal lines represent the CNA profile for individual CRPC samples (dark blue, homozygous loss; light blue, heterozygous loss; grey, no change; pink, copy number gain (CNA 3–4); red, copy number amplification (CNA > 4); white, insufficient data to determine CNA). Samples are clustered on the basis of *RNASEH2B* gene status. CNA frequencies for *RNASEH2B* and the *RB1–RNASEH2B* region without a copy number breakpoint are shown on the right.

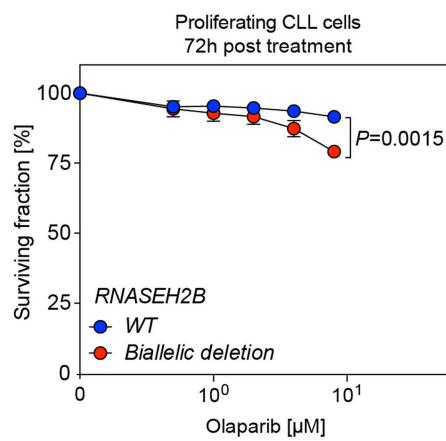
a



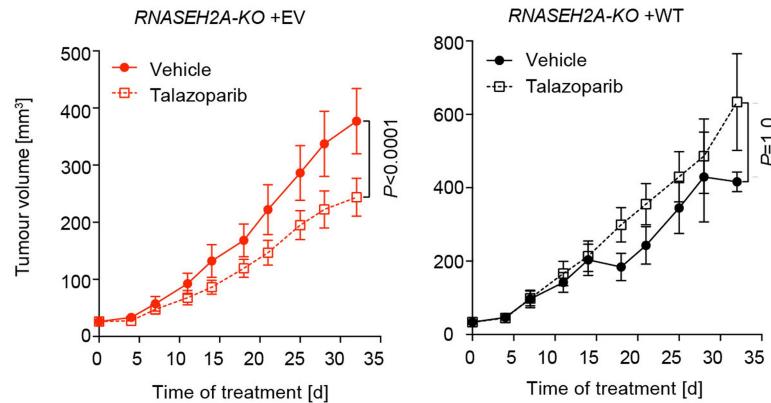
b



c



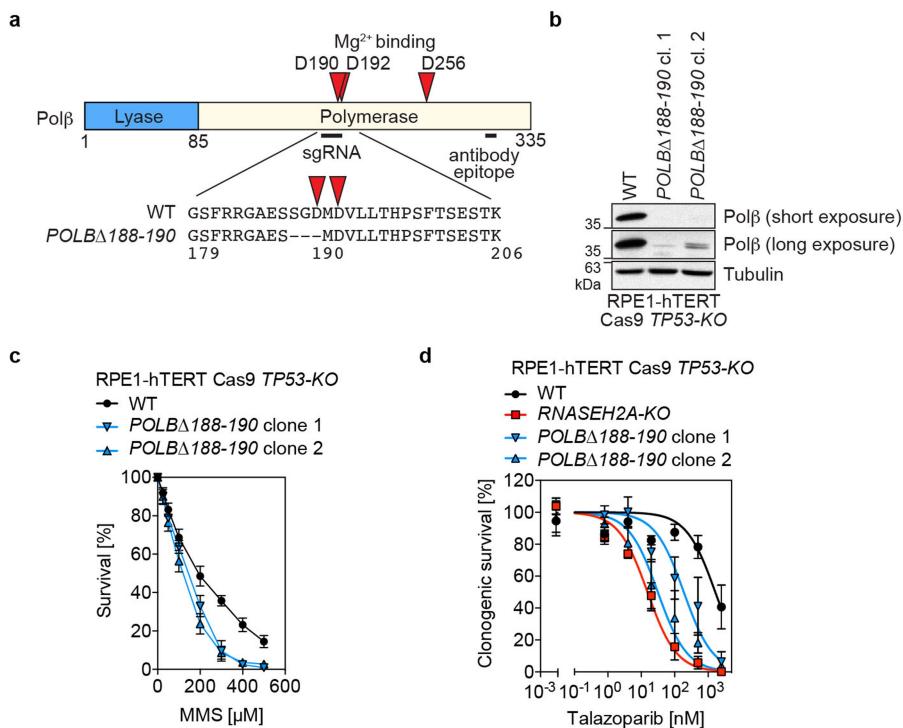
d



Extended Data Fig. 8 | PARPi sensitivity in RNase H2-deficient primary CLL cells and mouse xenograft tumours. This figure is related to Fig. 4. **a, b**, Proliferating cells, and not quiescent cells, are the major population of viable cells in ex vivo cultured primary CLL patient samples irrespective of treatment group. Quantification of absolute (a) and relative (b) quiescent and proliferating cell numbers as determined by FACS analysis of the primary CLL samples used in Fig. 4b, c. Wild-type RNASEH2B, $n=8$ individual samples; monoallelic deletion, $n=4$ individual samples; biallelic deletion, $n=9$ individual samples. Data are mean \pm s.d. from three technical replicates. FACS gating strategy for stimulated peripheral blood lymphocytes (PBLs) from CLL patients is shown in Supplementary Fig. 2. **c**, RNase H2-deficient primary CLL cells have reduced survival

when cultured with olaparib. Data are the mean of individual samples \pm s.e.m. ($n=3$ biologically independent CLL samples per group, each analysed in technical triplicates). P value from a two-way ANOVA.

d, Talazoparib selectively inhibits the growth of RNASEH2A^{KO} xenograft tumours. RNASEH2A^{KO} cells complemented either with empty vector or wild-type RNASEH2A were injected subcutaneously into bilateral flanks of CD-1 nude mice. Mice were randomized to either vehicle or talazoparib (0.333 mg kg^{-1}) treatment groups (eight animals per group) and tumour volumes were measured twice weekly. Data are mean \pm s.e.m. P values are from two-way ANOVA under the null hypothesis that talazoparib does not suppress the tumour growth.



Extended Data Fig. 9 | Rnase H2-deficient cells are more sensitive to PARPi than DNA polymerase β mutants. **a**, Schematic of the $POLB^{\Delta 188-190}$ CRISPR mutation. The Mg²⁺-coordinating aspartate residues (D190, D192 and D256, red triangles) are highlighted in the domain structure of the human DNA polymerase β protein. The sgRNA target site and antibody epitope are indicated by black lines. **b**, WCEs from parental RPE1-hTERT Cas9 $TP53^{KO}$ cells and two $POLB^{\Delta 188-190}$ clones were processed for immunoblotting with DNA polymerase β and tubulin (loading control) antibodies. Immunoblots are representative of two biologically independent experiments. **c**, The $POLB^{\Delta 188-190}$ mutation

impairs base excision repair. Wild-type or $POLB^{\Delta 188-190}$ RPE1-hTERT Cas9 $TP53^{KO}$ cells were exposed to different concentrations of MMS for 24 h, and then grown in drug-free medium for an additional 48 h. Cell viability was determined using the Cell Titer Glo assay. **d**, Sensitivity of wild-type, $RNASEH2A^{KO}$ and $POLB^{\Delta 188-190}$ RPE1-hTERT Cas9 $TP53^{KO}$ cells to indicated talazoparib concentrations in clonogenic survival assays. Data in **c**, **d** are mean \pm s.d. from three biologically independent experiments normalized to untreated cells. Solid lines show a nonlinear least-squares fit to a three-parameter dose-response model.

Extended Data Table 1 | Clinical and molecular characteristics of primary CLL samples

Clinical characteristics							Molecular characteristics					
Sample	Age	Sex	Binet stage	Time from diagnosis (Months)	Treatment	Time on treatment (Days)	Response to treatment	Cytogenetics (FISH)	<i>RNASEH2B</i> status ¹	<i>ATM</i> status ²	<i>TP53</i> status ³	<i>IgVH</i> status ⁴
CLL1	67	F	A	35	Pre-treatment	0	-	Trisomy 12	WT	WT	WT	M
CLL2	74	F	A	24	Pre-treatment	0	-	Normal	WT	WT	c.658_663del, c.849_850insC#	UM
CLL3	67	M	A	176	Ibrutinib	0	PRL	Normal	WT	WT	WT	UM
CLL4	68	M	A	49	Pre-treatment	0	-	Normal	WT	WT	WT	M
CLL5	76	M	A	49	Pre-treatment	0	-	N/A	WT	WT	WT	UM
CLL6	65	F	A	153	Pre-treatment	0	-	N/A	WT	WT	WT	M
CLL7	63	F	A	199	Fludarabine+Cyclophosphamide+Rituximab	37	CR	Trisomy 12	WT	WT	WT	UM
CLL8	39	M	B	80	Pre-treatment	0	-	Normal	WT	WT	WT	M
CLL9	80	F	A	33	Chlorambucil	83	PR	del(13q)	Monoallelic del	WT	WT	M
CLL10	57	F	A	136	Pre-treatment	0	-	del(13q)	Monoallelic del	WT	WT	M
CLL11	79	F	A	70	Bendamustine + rituximab	251	CR	N/A	Monoallelic del	WT	WT	M
CLL12	48	M	B	159	Ibrutinib	486	PR	N/A	Monoallelic del	WT	WT	UM
CLL13	62	F	A	203	Pre-treatment	0	-	N/A	Biallelic del	WT	WT	M
CLL14	63	M	A	27	Pre-treatment	0	-	del(13q)	Biallelic del	WT	WT	UM
CLL15	42	F	A	414	Bendamustine + rituximab +/- ibrutinib	120	SD	del(13q)	Biallelic del	WT	c.561A>G *	M
CLL16	84	F	A	19	Pre-treatment	0	-	N/A	Biallelic del	WT	WT	M
CLL17	72	F	A	153	Chlorambucil	63	PR	Trisomy 12, del(13q)	Biallelic del	WT	c.743G>A*	M
CLL18	79	F	A	36	Pre-treatment	0	-	del(13q)	Biallelic del	WT	WT	M
CLL19	48	F	B	8	Pre-treatment	0	-	del(17p), del(13q)	Biallelic del	WT	c.753_754insCC#	M
CLL20	70	F	B	10	Pre-treatment	0	-	del(13q)	Biallelic del	WT	WT	UM
CLL21	67	M	B	56	Pre-treatment	0	-	del(13q)	Biallelic del	WT	WT	UM

Data used in Fig. 4a, c, Extended Data Figs 7a, b, 8a, b. CLL samples grouped by *RNASEH2B* status. CR, complete response; F, female; M, male; N/A, not available; PRL, partial response with lymphocytosis; PR, partial response; SD, stable disease; dashes (-), not applicable.

¹Based on MLPA and CGH array. del, deleted.

²Intact *ATM* status confirmed by next-generation sequencing and/or functional assays.

³*TP53* status determined by sequencing. *, monoallelic *TP53* alteration; #, biallelic *TP53* alteration.

⁴Maturation status of CLL assessed by detection of hypermutation in immunoglobulin variable region heavy chain (*IgVH*); UM, unmutated (more than 98% sequence homology with germline sequence); M, mutated (less than 98% sequence homology with germline sequence).

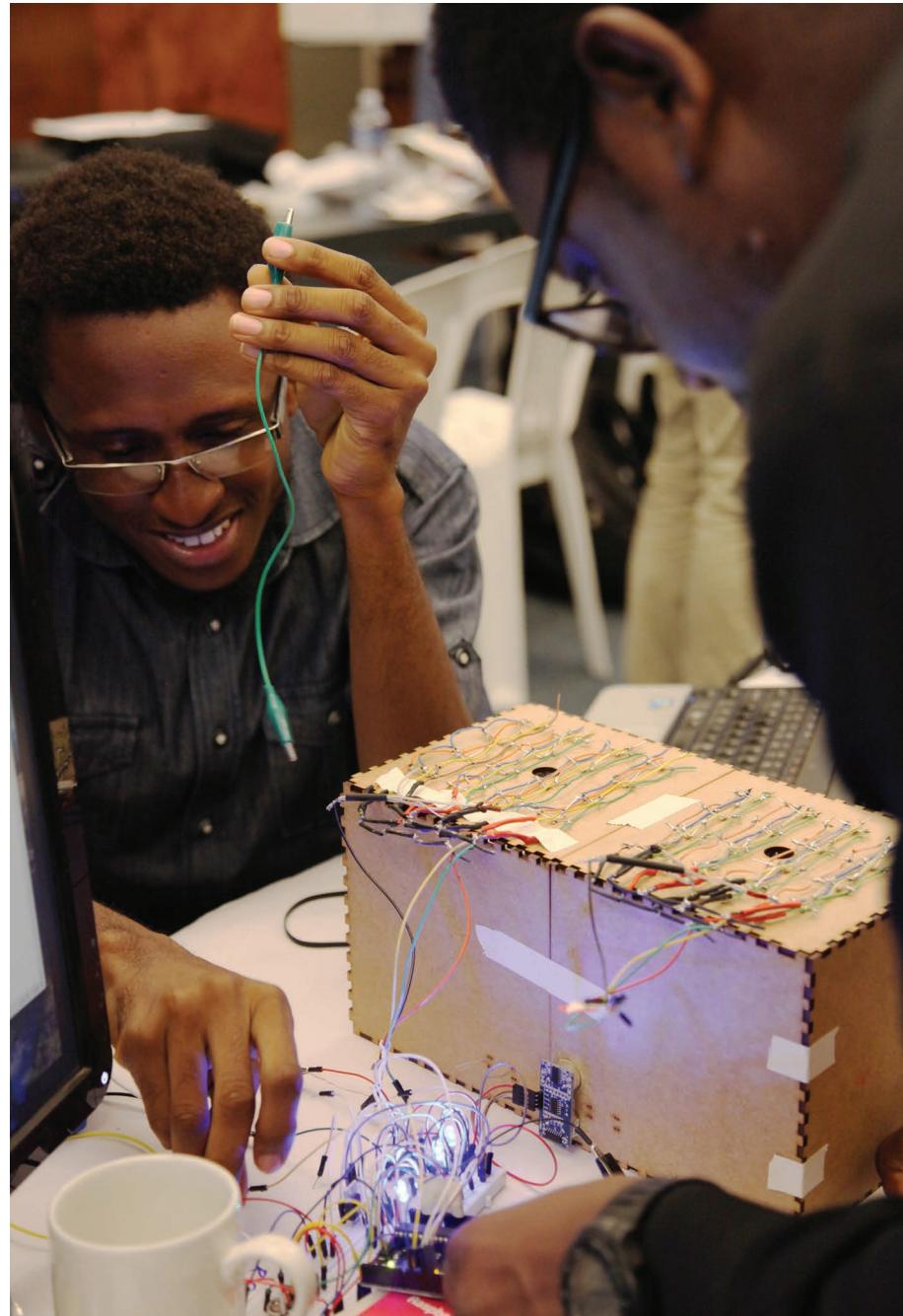
CAREERS

TRAINING Institutions must broaden their scope beyond academia, **p.293**

UNIVERSITIES Improve representation and pay for ethnic-minority women, **p.293**

 **NATUREJOBS** For the latest career listings and advice www.naturejobs.com

AGNIESZKA POKRYWKA/TREND IN AFRICA



Oluwaseun Faborode (right) and another scientist enjoy a workshop on how to make low-cost lab gear.

FRUGALITY

Labs on the cheap

How to create a champagne setup on a beer budget.

BY ELIE DOLGIN

Microbiologist Rebecca Shapiro faced a daunting task after starting a tenure-track job at the University of Guelph in Canada: building a laboratory from scratch, on a tight budget.

She inherited some equipment from retiring faculty members but much of it was broken — dangerously so in a few cases. In one instance, she almost burnt herself on a used heating block donated by a colleague, because the whole device became red-hot when she turned it on. “I was like, ‘OK, that’s going in the garbage,’” she recalls.

In June 2018, 6 months into her new job, Shapiro applied to the Canadian and Ontario governments for a Can\$200,000 (US\$152,000) infrastructure grant. The funds would help her to buy most of her equipment, including a Can\$40,000 plate reader, a Can\$40,000 shaking incubator and a Can\$20,000 ultralow-temperature freezer. But she has also turned to online auction sites such as eBay, BidSpotter and BioSurplus to find deals on functioning heating blocks and other smaller items to get her lab operational.

“My lab thinks I’m weird as I’m constantly on my phone going, ‘Oh no, we’ve been out-bid,’” Shapiro says. But the exercise paid off. She estimates that, so far, she’s bought most of her gear — pipettes, vortexes, centrifuges, pH metres, hot plates and more — at a markdown of 60–90% from typical catalogue prices.

Most new principal investigators (PIs) face budget constraints and need to be shrewd about securing lab equipment. They might have start-up packages that, in real terms, are very large sums, but just a few pieces of high-end kit can quickly eat up those funds. Meanwhile, science labs in countries with fewer resources might operate on a shoestring, and new PIs in the developing world often struggle to procure even the most basic supplies.

Fortunately, money-saving deals and inexpensive workarounds abound. Scientists just need to know where to look for them.

It takes extra work to find bargains: some PIs even resort to begging for and borrowing old equipment from other labs, or they work out ways to share things. Yet even the most cash-strapped new lab leaders can usually get their research spaces up and running quickly. They can scour discount online sites, check out businesses such as hardware stores or restaurants that might have similar or identical equipment to sell, and hold off on buying immediately from vendors, who might offer a lower price ▶

► later on if the item has not already been sold. As an added bonus, says Shapiro, fitting out a lab on the cheap can provide a personal sense of accomplishment. "It can actually be really fun," she says. "I recommend it."

ECONOMICAL SHARING

One way to save on big-ticket items is to avoid purchasing things that are available already for communal use at core or shared-research laboratories — facilities in which scientists can either book time on state-of-the-art equipment or pay staff to perform technically demanding experiments on their behalf.

At Imperial College London, for example, the department of materials offers a range of top-of-the-line technologies, including electron microscopes and focused-ion-beam instruments — both things that condensed-matter physicist Ben Britton knew he needed for his research into the nature of metals used in the aerospace and energy industries. When Britton was considering whether to take a faculty position in the department, he made sure the job would give him ample time to access these machines affordably.

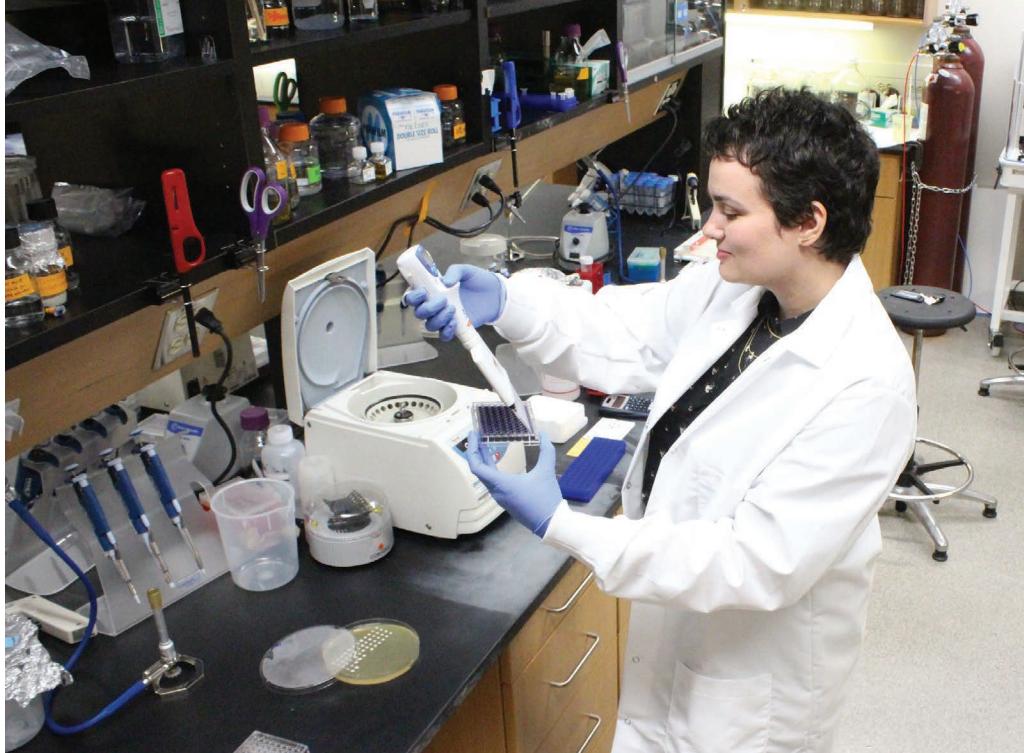
Before accepting the post, he negotiated unlimited free access to the equipment until he secured external funding. Britton estimates that the arrangement saved him an extra £20,000 (US\$26,500) annually in his first few years as an independent leader, a sum he used instead for computing equipment.

Thinking beyond core research facilities, chemist Paul Bracher recommends setting up less-formal arrangements with nearby labs. This can save researchers from having to buy equipment that's integral for their work, but that doesn't get used every day. Bracher, who started a lab five years ago at Saint Louis University in Missouri, studies the origin of life. His research involves some chemical synthesis — but not a lot. So he managed to avoid buying a \$5,000 rotary evaporator, a device that removes solvents from samples, by arranging to borrow one when necessary from a colleague.

For Bracher, each dollar saved was another dollar available to buy something else for his lab. To work out what he needed, he created a spreadsheet and sought advice from other junior faculty members who had been in a similar position, and from equipment vendors who had worked with other new PIs.

Among the spreadsheet entries were everyday lab items such as hammers and paraffin film — things that typically include huge markups when bought from lab supplies companies. Instead, Bracher popped down to his local hardware store and shopped online. A \$70 hammer was available at one-tenth of the price. On Amazon, a \$100 roll of paraffin film sells for around \$25.

But it's not just household goods that are sold by alternative vendors. Sometimes it's worth thinking across industries. Bracher's research, for example, often involves setting



Microbiologist Rebecca Shapiro says fitting out your own lab on the cheap "can actually be really fun".

up reactions in humidity-controlled chambers that are designed to mimic the conditions in primordial times. A good-sized chamber can run into the tens of thousands of dollars — much more than Bracher had budgeted for. Restaurants and catering businesses, however, use practically the same item — known as a 'holding cabinet' — for keeping cooked food hot and ready to eat.

Those run for about \$5,000 apiece. Bracher picked up three of them. And although the food-grade cabinets might not be quite as precisely fine-tuned as the lab-grade ones, they're more than adequate for his research purposes.

Bracher also used crafty negotiating tactics for items he could get only from specialized lab suppliers. For instance, he knew he wanted an expensive type of mass spectrometer for analysing trace metals, but he didn't need it right away. So he put in a few low offers and waited for vendors to meet his price. Months later, a sales representative called. He was trying to meet a manager's quota, he explained, because it was the end of a fiscal quarter, and he could now offer the instrument for almost as low a price as Bracher had initially offered. "It often helps to wait," he says.

But haggling to the point of annoyance can backfire, warns Kevin Ryan, who until he retired in April co-owned and operated W. Nuhsbaum, a microscope dealership headquartered in McHenry, Illinois. Purchasing an expensive microscope or some other costly piece of lab gear might seem a bit like buying a car, he says. But whereas any mechanic can change a car's oil and rotate its tyres, the sales representative who sells a large lab instrument to a PI will be the researcher's point of contact for servicing and upgrades for years to come.

"It's a fine line with discounting," Ryan says. "You don't want to be too hard on the salesperson because you want to build a long-lasting relationship."

Besides, there's often no need for aggressive negotiation tactics; most suppliers will accommodate reasonable requests to secure the business of a newly hired faculty member, notes Lisa Witte, president of Fisher Scientific, a lab-supply company in Pittsburgh, Pennsylvania. That's because early-career investigators have decades of purchasing ahead of them, and vendors want to build brand loyalty early on. As Witte points out: "We want to earn that repeat business."

That's why Fisher Scientific created its New Lab Start-Up Program, and why other suppliers — including MilliporeSigma in Burlington, Massachusetts, and VWR in Radnor, Pennsylvania — offer something similar. Fisher Scientific's scheme works like a coupon book for lab supplies, with 100 money-saving offers across a broad range of products for new PIs. "We really try to help the PI stretch those precious research dollars as far as they can," Witte says. The more a PI spends, the more they can save, she adds, "because if it's one big bulk order, we're generally able to give additional discounts".

MONEY PROBLEMS

Even with heavy discounting, however, some researchers still can't afford even the most basic labware. Plant biologist Muvari Tjiurutue completed a PhD at the University of Massachusetts Amherst before taking a job in 2016 at the University of Namibia in Windhoek. There, she had no shakers or stirrers, let alone any of the more sophisticated analytical instruments that she had come to rely on in graduate school. "I haven't actually been able to do my research because of lack of equipment," Tjiurutue says. "It is very frustrating."

Seeding Labs, a non-profit organization in Boston, Massachusetts, came to her rescue. It collects millions of dollars worth of surplus equipment from Western universities,

companies and government agencies. It then offers this inventory to institutions in low- and middle-income countries that show potential to advance cutting-edge research — if only they had the gear.

Tjiurutue applied to Seeding last July, asking for equipment that she estimates was worth more than 5 million Namibian dollars (US\$365,000). A shipment full of chromatographers and basic lab staples is now due to arrive before the year's end. Tjiurutue's department only had to pay N\$33,500 to defray some of the costs of obtaining, handling and shipping the equipment.

Rupika Delgoda, a chemist at the University of the West Indies in Mona, Jamaica, received her own delivery of lab goods from Seeding Labs in October 2017, but not before struggling for years to launch the university's Natural Products Institute. "We had to build counter-tops and start from scratch," says Delgoda, who heads the institute.

Her strategy was to ask former colleagues for unused equipment. Her ex-lab mates from the universities of Oxford and Leicester, UK, where Delgoda had trained, came through with boxes of free gear. "They were happy to know they found a good home," Delgoda says.

Others also find ways to get by without buying anything. Evolutionary geneticist Santiago Castillo Ramírez returned from a postdoc in the United Kingdom to start a lab at the Center for Genomic Sciences in Cuernavaca, part of the National Autonomous University of Mexico. Because he had a minimal start-up package that barely covered the cost of a couple of computers, he decided to form collaborations.

Initially, he partnered with teams in Germany and the United States, working with genomic data sets they'd previously amassed on sexually transmitted infections and tick-borne pathogens. Then he joined forces with Miguel Cevallos, an experimental microbiologist also working at the Center for Genomic Sciences. Together, they established a research programme studying the rise of new kinds of drug-resistant bacteria in Mexican hospitals.

Cevallos does the lab work and Castillo Ramírez sticks to the genomic analysis. "It was a win-win situation for both of us," Castillo Ramírez says.

MAKE DO

Scientists with access to a 3D printer and a bit of engineering know-how now also have the opportunity to make their own lab equipment — with detailed assembly instructions and open-source software available through online repositories. "It's a very efficient way

of getting things done quickly to a fairly high standard," says Tom Baden, a neuroscientist at the University of Sussex in Brighton, UK.

Baden's lab uses an Ultimaker 2, a compact desktop machine that costs around £1,800 new, but alternatives exist for a few hundred pounds. And although it takes a little longer to build this kind of gear in-house, he notes, the material and electronics that go into 3D-printed labware typically cost a fraction of what commercial platforms would charge.

Last year, for example, Baden detailed the design blueprint for making a pressure ejection system for precisely delivering minute volumes of liquid using off-the-shelf components and 3D-printed parts (C. J. Forman *et al. Sci. Rep.* 7, 2188; 2017). It cost him around £400 in materials, he says; commercial models are at least five times as much. His group has also developed and produced other custom-built gear. "Anything mechanical that doesn't need to be micrometre-precise tends to be 3D-printed in our lab," Baden says.

Through a non-profit organization that he co-founded, called TReND in Africa, Baden now runs workshops in Ethiopia, Uganda, South Africa and elsewhere to train scientists in this kind of do-it-yourself approach to low-cost labware. Oluwaseun Faborode, a neurophysiologist at the University of Ibadan, Nigeria, attended one of those courses and quickly put his newfound skills to use, programming a microcontroller to aid him in studying rodent models of depression. "It's just a basic timing indicator," he says — but it helps to reduce the chance of human error and investigator interference in the experiments. By building his own equipment, Faborode says, "I'm upping my game."

But relying on homemade gear or second-hand lab supplies has its downsides. There's a chance of something being faulty, and there's no money-back guarantee from the manufacturer. On the other hand, "you don't have an unlimited pot of money", notes Bracher, so one must be strategic about where to cut corners and when to pay full price.

Shapiro draws the line at her plate readers, instruments she uses to measure simultaneously the growth of dozens or hundreds of yeast strains to probe drug resistance. "You're running some risk of the equipment not working very well, and it's not under warranty," she says. "I'm not willing to run that risk on some \$40,000 sensitive piece of analytical equipment, but I am willing to run that risk on a basic centrifuge or vortex."

One scale that she bought on Bidspotter came without its charging cord. And the electronic multichannel pipette from eBay needed to be recalibrated. But Shapiro got both items working eventually.

And the added hassle? "It's easily worth it," she says. ■

Elie Dolgin is a freelance writer in Somerville, Massachusetts.

TRAINING

Broaden careers advice

A consortium of European research universities is calling for its members to support junior scientists' efforts to pursue non-academic careers. The League of European Research Universities (LERU), which represents 23 institutions, says in a June publication that universities, supervisors and principal investigators (PIs) overemphasize an academic career path, even though more than 60% of all European research jobs lie outside academia (see go.nature.com/leru). The report cites a "critical need" for training and support programmes to help graduate students and postdoctoral researchers prepare for a wide possibility of career paths. It also points out that some universities have already launched such initiatives. Several, for example, have programmes aimed at helping PhD students and postdocs to launch their own businesses. LERU recommends that PIs and supervisors counter the common perception that any job outside academia counts as a failure. To further help young researchers gain experience and independence, LERU calls on the European Commission and other funding bodies to support research in which postdocs serve as project leaders.

EQUALITY

Call for standards

US universities should make their hiring and pay practices fairer for women from ethnic minorities who are faculty members, administrators or other academic professionals, says an association of human-resources executives working in higher education. In a report published in May (see go.nature.com/cupa_hr), the College and University Professional Association for Human Resources (CUPA-HR) in Knoxville, Tennessee, finds that women from ethnic minorities in faculty and administrative positions earn less than 87% of their white male counterparts' salaries, and that their numbers are disproportionately low in those positions. The association advises institutions to evaluate recruitment and pay practices for all job categories, and to compare their salary data with those of peer institutions to establish fair market rates for employees in minority groups. CUPA-HR also encourages universities to consider suggestions for improving promotion potential for female employees from ethnic minorities. The report is based on surveys carried out in 2016–17 of nearly 560,000 employees at more than 1,100 public and private institutions.

YOUR FACE

The price of success.

BY GRACE TANG

“Welcome home, Mendel.” I greeted my DNA Scraper unit as it returned home for the day with its finds. I realized I was talking to my robots — probably a consequence of spending the past year living and working alone.

I picked Mendel up from the ground, popped its lid open, and dumped the contents of the sample compartment into the desktop gene-sequencing apparatus. Cigarette butts, small vials of excrement scrapings, more needles than I could count... The city ought to be paying me at the rate Mendel was cleaning up the homeless encampments for them.

I microwaved yesterday's leftovers while the analysis ran. For more than a year since I'd quit the force, this had been my daily routine. Too hungry and impatient to wait for the first dish to be ready, I grabbed another box straight from the fridge and ate the contents cold. I sighed. I just needed to be patient till I hit the jackpot.

The microwave's beeping sounded off — I was wondering if it had received a software update when I realized the unfamiliar chime was in fact coming from the gene sequencer. Slightly concerned, I went to investigate, wondering if the warranty was still valid.

I nearly choked on my cold day-old mushu pork.

For the first time since I'd acquired this unit, its screen displayed a face — your face, or at least the best guess the algorithm had generated, based on the DNA extracted from one of Mendel's samples. You were a match.

The algorithm tweaked the features of the 3D model as its confidence in aspects of your appearance grew.

You were of majority Caucasian descent, with a wide, flat nose, thin lips and brown hair (the 95% confidence interval was between dirty blonde and deep chestnut). Telomeric analysis placed your biological age at around 40.

But what caught my attention were your eyes — amber, fading to green around the edges (96% confidence), staring back at me from the screen.

Once upon a time, the world had had a

diversity of eye colours. Blue, green, violet, you name it. But over time, as the people of the world overcame the barriers that separated them, genes intermixed, and dominant genes turned the globe's eyes into a monotonous shade of brown, rendering any other eye colours a rarity.

found you slumped in an alley, drifting in and out of consciousness, surrounded by used needles and puddles of piss and vomit. Your frazzled hair stuck out of your head like an unkempt jungle, and your cheekbones jutted out of your malnourished face, but otherwise, you were a spitting image of the face that had stared at me from my computer back home. Your arms were covered in sores. Perhaps it was better, what they were going to do to you.

I held my nose with one hand to block the stench of old urine and puke, and swabbed you with my other hand — you barely noticed. I ran your sample through my portable kit.

The green light confirmed a match to the DNA from the needle. Somewhat reluctantly, I called the number.

It was all happening so fast. An ambulance rolled down the alley and pulled up beside us.

“This him?” The ‘paramedic’ asked. I nodded.

He grinned wordlessly — tonight we would both be eating well. I caught a glimpse of your green eyes as they loaded you on a stretcher and put you in the back of the ambulance.

Somewhere, today, our client would be jumping several spots up to the top of the organ-donation waiting list, because, miraculously, a close genetic match had been found (totally without the help of a retired inspector using self-refurbished police equipment smuggled from the forgotten depths of the electronics-recycling centre). Fortunately, the donor had died of an overdose before the ambulance could get him to the hospital (despite the best efforts of the paramedics, who absolutely did not inject him with even more of the substance that was killing him).

No one would miss one more nameless addict off the streets.

That night, I received the password to an untraceable wallet containing more money than I had ever seen in one place. If you had asked me a year ago what I thought I'd feel when I finally completed this assignment, I would not have guessed that it would feel like this. I ordered my first decent dinner in a year, powered down Mendel and Argos, and tried to forget your face. ■

After obtaining a PhD in psychology from Stanford University, Grace Tang now works as a data scientist in the tech industry.



To have two recessive genes for light eyes, you must have been something special. I thought of the royal family and their equally purebred corgis. Maybe you were royalty...

But probably not. Your sample had come from a needle.

Punching in instructions to analyse the non-genetic material in your sample, I confirmed the presence of heroin. Thankfully, the tests for HIV, and hepatitis B and C came out negative, otherwise my client would not have been pleased. I hoped you were still alive.

The next day, I dusted off Argos, a much more sophisticated Seeker model. I inserted an amplified vial of your DNA and synced the results of the predictive algorithm with the coordinates of where your first sample was found, then set Argos off into the world.

I had waited more than a year for Mendel to find a sample with sufficient genetic overlap with one of my clients, but it wasn't even a full day before Argos found you. I guess addicts did not travel far from the things they left behind. I had not expected my long-awaited triumph to be laced with regret.

I set out to the coordinates Argos sent, and

© NATURE.COM

Follow Futures:

Twitter: @NatureFutures

Facebook: go.nature.com/mtoodm